

universität ulm

Individual Differences in Knowledge: Examining Test Modes and Putting the Nomological Net Into Perspective

CUMULATIVE DISSERTATION

for the degree of

DOCTOR RERUM NATURALIUM

(Dr. rer. nat.)

presented to the Faculty of Engineering,

Computer Science, and Psychology at Ulm University, Ulm, Germany.

by

Benjamin Goecke

Born in Ludwigsburg

2023

Institute for Psychology and Education

Department of Individual Differences and Psychological Assessment

Department Head: Prof. Dr. Oliver Wilhelm

Day of Promotion: 13th July 2023

Dean: Prof. Dr. Anke Huckauf

1. Reviewer: Prof. Dr. Oliver Wilhelm

2. Reviewer: Prof. Dr. Aljoscha Neubauer

3. Reviewer: Prof. Dr. Matthias Ziegler

Acknowledgments

Here it is, my dissertation. Along with this text comes a deep gratitude for my position in life. In this context, I would like to extend my deepest gratitude to everyone who has supported me throughout this journey.

Firstly, I want to express my sincerest appreciation to my supervisor, Oliver Wilhelm, for his endless support, guidance, and invaluable insight into the field of individual differences research. Without his mentorship, this dissertation or any work apart from it during my studies would not have been possible.

Thank you, Oliver.

I am also immensely grateful for my colleagues whose expertise, dedication, and collaborative spirit contributed significantly to the success of this endeavor. Thank you, Selina Weiss, Johanna Hartung, Mattis Geiger, Luc Watrin, Elisa Altgassen, Franziska Rees, Ulrich Schroeders, Diana Steger, and Annette Kniep. I am honored to have worked alongside you all and feel fortunate to have made such strong connections with all of you. Naturally, I am looking forward to keeping doing research with you all. Or, we just hang out.

Thank you.

Whenever I refer to "I" in this dissertation, I acknowledge the contributions of all my co-authors to my studies as well. Please know that I have learned a lot from all of you and your extraordinary expertise.

Thank you.

Also, I would like to thank Prof. Matthias Ziegler and Prof. Aljoscha Neubauer for agreeing to review this dissertation. I do appreciate the time and effort that you will dedicate to reading my work.

Thank you.

To my partner, Selina: words cannot express how grateful I am for your unwavering support, love, and encouragement throughout this journey from the beginning. Your patience and understanding during the most challenging moments of this process have meant the world to me, and I could not be happier.

Thank you.

To my little daughter, Soley: you have been the sunshine of my life ever since you arrived in this utterly strange world. It had to be you, and I could not have wished for a better baby companion during the last six months of this dissertation. You are perfect, and I will always be there for you in endless affection.

Thank you.

Lastly, I want to express my heartfelt appreciation to my parents, Barbara and Tom, whose love, encouragement, and ever-lasting support have been a foundation of my academic and personal success. I am deeply grateful for their constant support and unwavering belief in me, no matter what I decided to take on.

Thank you.

As I reflect on this journey and look forward to the next chapter of my life, I am humbled by the support and encouragement that has helped me reach this point.

Thank you to all who have been a part of this journey.

Dedicated to my little daughter, Soley.

Contents

I. Prologue

Introduction	I – 2
Closing in on (Declarative Knowledge)	I – 6
Measurement of Declarative Knowledge	I – 10
Declarative Knowledge as a Vault of Information	I – 15
Overview of the Dissertation Manuscripts	I – 18
References	I – 21

II. Manuscript I

Testing Competing Claims About Overclaiming

Introduction	II – 1
Scoring methods of overclaiming: Raw responses vs. signal detection indices	II – 2
What drives overclaiming behavior?	II – 2
Research aims	II – 4
Method	II – 4
Measures	II – 4
Statistical Analysis	II – 5
Results	II – 6
Measurement Models	II – 6
Overclaiming as a self-enhancement tendency	II – 6
Overclaiming as an index of cognitive ability	II – 7
Overclaiming as a form of creative engagement	II – 8
Discussion	II – 8
Overclaiming as a self-enhancement tendency	II – 8
Overclaiming as a cognitive bias	II – 9
Overclaiming as a proxy for cognitive abilities	II – 9
Conclusion	II – 9
References	II – 10

III. Manuscript II

The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children

Introduction III – 1
Scoring of OCQs III – 1
What causes OC behavior in children? III – 2
Purpose of the current study III – 3
Method III – 3
Participants and procedure III – 3
Measures III – 4
Statistical Analysis III – 4
Results III – 5
Measurement Models III – 5
Discussion III – 7
Perspectives on overclaiming III – 8
On the Use of OCQs III – 8
Where do we go from here? III – 9
Conclusion III – 10
References III – 10

IV. Manuscript III

Stop Worrying About Multiple Choice: Fact Knowledge Does Not Change with Response Format

Introduction IV – 1
Declarative Knowledge as an Indicator of Crystallized Intelligence IV – 2
Assessment Methods of Declarative Knowledge IV – 2
MC Format Items vs. CR Format Items IV – 3
Cognitive Processes Underlying Different Response Formats IV – 4
The Present Studies IV – 5
Methods and Materials: Study 1 IV – 6
Participants and Procedure IV – 6
Measures IV – 6
Data Preparation IV – 7
Statistical Analysis IV – 8
Results: Study 1 IV – 8
Discussion: Study 1 IV – 9
Methods and Materials: Study 2 IV – 10
Measures IV – 11
Data Preparation IV – 11
Statistical Analysis IV – 11
Results: Study 2 IV – 12
Discussion: Study 2 IV – 14
General Discussion IV – 15
Response Formats as Means to an End IV – 15
Recognition, Recall, or What to Study Next IV – 16

Limitations IV – 18
Conclusions IV – 18
References IV – 19

V. Manuscript IV
Driving Factors of Individual Differences in Broad Retrieval Ability: Gr *Is More Than the Sum of Its Parts*
Introduction V – 4
Dimensionality of Gr V – 4
Cognitive Processes Underlying Gr V – 6
Empirical Findings Regarding the Nomological Net of Broad Retrieval Ability V – 11
The Present Study V – 13
Method V – 14
Participants and Procedure V – 14
Measures V – 15
Data Treatment V – 21
Statistical Analysis V – 21
Results V – 22
Are We Able to Distinguish Between Short and Long Periods of Retrieval in the Verbal Domain? V – 22
Measurement Model of Gr Including Figural Fluency V – 26
Predictors of Individual Differences in Gr V – 27
Discussion V – 31
The Measurement Model of Gr V – 32
The Nomological Net of Gr V – 33
Desiderata for Future Research V – 37
Conclusion V – 39
References V – 40

VI. Epilogue
Summary VI – 2
Manuscript I: Testing Competing Claims About Overclaiming VI – 2
Manuscript II: The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children VI – 4
Manuscript III: Stop Worrying About Multiple-Choice: Knowledge Does Not Change with Response Format VI – 6
Manuscript IV: Driving Factors of Individual Differences in Broad Retrieval Ability: *Gr* Is More Than the Sum of Its Parts VI – 7
Looking Towards the Empirical Findings and Beyond VI – 9
Question 1: How Is Self-Reported Knowledge to Be Placed Within The Existing Ability Structure? VI – 10

Question 2: How Do Different Response Formats Affect the Measurement of Declarative Knowledge? VI – 14
Question 3: Which Cognitive Abilities Explain Individual Differences In Broad Retrieval Ability? VI – 15
Being Wise About Studying Knowledge and Related Constructs VI – 18
Overclaiming About Overclaiming VI – 18
The Idiosyncratic Structure of Knowledge and Semantic Memory VI – 20
Conclusion VI – 23
References VI – 24

VII. Appendix
Supplementary Material Manuscript I VII – 1
Supplementary Material Manuscript II VII – 17
Supplementary Material Manuscript III VII – 29
Supplementary Material Manuscript IV VII – 43

List of Tables

II – Table 1	Synopsis of Theoretical Perspectives on Overclaiming and Corresponding Predictions	II – 3
II – Table 2	Descriptive Statistics for Familiarity Ratings, Including Means, Standard Deviations, Skews, and Kurtosis for Each Parcel	II – 5
II – Table 3	Measurement Models for Overclaiming, Self-Reported Knowledge, Crystallized and Fluid Intelligence, and Creativity	II – 6
III – Table 1	Descriptive Statistics for Familiarity Ratings, Including the Means, Standard Deviations, Skewness, and Kurtosis For Each OCQ Parcel	III – 5
III – Table 2	Model Fit for the Measurement Models of all Scales	III – 5
III – Table 3	Parameter Estimates of Strict Invariance Model Depicting OC, SRK, and Gc	III – 7
IV – Table 1	Design of the Studies	IV – 7
IV – Table 2	Logistic Regression Model With Standardized Regression Weights and Odds Ratios (OR)	IV – 9
IV – Table 3	Competing Measurement Models	IV – 13
V – Table 1	Four Component Model of Gr (Rosen & Engle, 1997) and Associated Cognitive Ability Constructs	V – 8
V – Table 2	Indicators of Broad Retrieval Ability (verbal)	V – 16
V – Table 3	Fit Indices of Measurement Models of Verbal Gr Accounting For Short and Long Retrieval Time Periods	V – 24
V – Table 4	Loadings of Indicators of Cognitive Covariates Measurement Model	V – 29

Appendix Manuscript I

VII – SM Table 1	Descriptive Statistics of All OCQ Items	VII – 1
VII – SM Table 2	Correlation Matrix of All Variables	VII – 7
VII – SM Table 3	Regression Models With d' (OC Accuracy) and c (OC Bias) As Predictors	VII – 11

Appendix Manuscript II

VII – SM Table 1	Descriptive Statistics of All OCQ Items	VII – 18
VII – SM Table 2	Measurement Models for All Scales	VII – 22
VII – SM Table 3	Measurement Invariance Testing for a Correlated Factor Model with OC, SRK, and Gc	VII – 23
VII – SM Table 4	Regression Models With d' (OC Accuracy) and c (OC Bias) As Predictors	VII – 26

Appendix Manuscript III

VII – SM Table 1	Descriptive Statistics of All Gc Items	VII – 30
VII – SM Table 2	Odds Ratios for All Items (Study 1)	VII – 32
VII – SM Table 3	Mean Differences Between Test Scores per Response Format for the Single Item Sets in Study 1	VII – 35
VII – SM Table 4	Means, Standard Deviations, Skew, Kurtosis, and Bivariate Correlations for all Indicators Used in the Measurement Models of Study 2	VII – 36
VII – SM Table 5	Measurement Models of the Response Formats	VII – 37

Appendix Manuscript IV

VII – SM Table 1	Full Manifest Correlation Matrix of All Indicators with Means and Standard Deviations	VII – 44

List of Figures

II – Figure 1	Structural Model of Overclaiming (OC) and Self-Reported Knowledge (SRK)	II – 6
II – Figure 2	Structural Model of Overclaiming (OC) and Self-Reported Knowledge (SRK), Honesty-Humility, and Faking Ability	II – 7
II – Figure 3	Structural Model of Overclaiming (OC) and Self-Reported Knowledge (SRK), and Knowledge (Gc), and Reasoning Ability (Gf)	II – 7
II – Figure 4	Structural Model of Overclaiming (OC) and Self-Reported Knowledge (SRK), and Divergent Thinking (DT), Originality and Openness	II – 8
III – Figure 1	Structural Model of OC and SRK	III – 6
III – Figure 2	Structural Model of OC, SRK, Openness to Experience, and Honesty-Humility	III – 6
III – Figure 3	Structural Model of OC, SRK, Gc, Gf, and Fluency	III – 6
III – Figure 4	Structural Model of OC, SRK, and Epistemic Curiosity	III – 7
IV – Figure 1	Mean Difficulties of Response Formats For Each Items Set	IV – 8
IV – Figure 2	Correlated Factors Model of the Response Formats	IV – 12
IV – Figure 3	Higher-Order Model of the Broad Knowledge Domains Across Response Formats With Defocusing as Predictor	IV – 12
V – Figure 1	Conceptual Measurement Models of Verbal Gr Accounting For Short and Long Retrieval Time Periods	V – 23
V – Figure 2	Higher-Order Measurement Model of Gr	V – 27
V – Figure 3	Measurement Model of Cognitive Covariates	V – 28
V – Figure 4	Latent Regression Model with Cognitive Covariates Predicting a Higher-Order Factor of Gr	V – 31

Appendix Manuscript I

VII – SM Figure 1	Correlation Plot of OC Accuracy and OC Bias Scores For All Four Cut-Offs	VII – 6

VII – SM Figure 2 Confirmatory Factor Model of Faking Ability and Knowledge (Gc) VII – 12
VII – SM Figure 3 Structural Model of Overclaiming (OC) and Self-Reported Knowledge (SRK) VII – 13
VII – SM Figure 4 Structural Model of Overclaiming (OC) and Self-Reported Knowledge (SRK), Honesty-Humility, and Faking Ability VII – 14
VII – SM Figure 5 Structural Model of Overclaiming (OC) and Self-Reported Knowledge (SRK), and Knowledge (Gc), and Reasoning Ability (Gf) VII – 15
VII – SM Figure 6 Structural Model of Overclaiming (OC) and Self-Reported Knowledge (SRK), and Divergent Thinking (DT), Originality and Openness VII – 16

Appendix Manuscript II
VII – SM Figure 1 Correlation Plot of OC Accuracy and OC Bias Scores For All Four Cut-Offs Across VII – 25

Appendix Manuscript III
VII – SM Figure 1 Correlated Factors of the Response Formats with Defocusing VII – 38
VII – SM Figure 2 A General Factor Model with Defocusing VII – 39
VII – SM Figure 3 Correlated Factors of the Broad Knowledge Domains Across Response Formats VII – 40
VII – SM Figure 4 Higher-Order Model of the Broad Knowledge Domains Across Response Formats VII – 41

Appendix Manuscript IV
VII – SM Figure 1 Measurement Model of Gr with Three Correlated Factors and Typing Speed as a Predictor VII – 48
VII – SM Figure 2 Latent Regression Model of Gr with Typing Speed as a Predictor VII – 49

Abstract

Declarative knowledge is a pivotal ingredient of crystallized intelligence (Gc). Although studying individual differences of knowledge has received considerable attention over the last two decades, misconceptions about both test modes and cognitive underpinnings of Gc are still prevalent. These shortcomings are in part due to a lack of multivariate studies pooling competing measures of declarative knowledge and, therefore, a limited theoretical perspective regarding antecedents and consequences of individual differences in knowledge.

This dissertation addresses these shortcomings by studying individual differences in declarative knowledge from broad theoretical and methodological perspectives. This includes the study of overclaiming and its nomological net as a salient bias of self-reported knowledge in adults and children (Manuscripts I and II), the study of how the choice of response format affects the measurement of maximal performance measures of declarative knowledge (Manuscript III), and a fourth larger study that places declarative knowledge and additional cognitive abilities into the nomological net of broad retrieval abilities (Gr; Manuscript IV).

In the first two manuscripts, I considered self-reports of knowledge as a possible test mode of declarative knowledge and examined relations with overclaiming (i.e., the exaggeration of knowledge). Within both multivariate studies, I examined the nomological net of self-reported knowledge and overclaiming with adults ($N = 298$) and children ($N = 897$). Both studies show that self-reported knowledge is substantially biased by overclaiming and only insufficiently associated with knowledge measures of maximal performance. I

summarize with both recommendations for measuring declarative knowledge and with ideas for future studies on overclaiming.

In the third manuscript, I report two studies (N = 142 and N = 300) that focus on methodological considerations regarding competing response formats (multiple-choice vs. open formats) for measuring declarative knowledge. Even though the use of different response formats affects the mean structure of a knowledge test, the results of latent variable analyses suggest that declarative knowledge does not change as a function of the response format.

In the fourth manuscript, I present a multivariate study (N = 331) of Gr, which posits another important factor of human cognitive abilities. I examine the construct validity of Gr given various experimentally manipulated conditions and juxtapose the construct with a number of relevant covariates, including a broad measurement of Gc. The results show that Gc is incrementally predictive above and beyond other cognitive abilities towards Gr. We discuss the reciprocal relationship between Gc and Gr. Our results endorse the position that Gr is more than a linear combination of its predictors.

Throughout the dissertation, I associate the main results of the four manuscripts, which add to the previous literature by critically testing theoretical accounts of the constructs studied, by providing sound empirical data from large multivariate studies, and by paving the way for future research efforts. The dissertation highlights the cumulative nature of research in the field and concludes with guidance regarding extensions and suggestions for studying Gc, related variables, and their cognitive underpinnings.

I Prologue

Introduction

"Of course knowledge counts."

- Earl Hunt (2000, p. 126)

One might argue that knowledge counts because knowing is better than not knowing. From an epistemological point of view, *knowledge* can be deemed as the Holy Grail that entails human existence (Steup & Neta, 2005) and from a sociological-evolutionary view as a cumulative good of human society that is essential for our survival (c.f., cumulative cultural evolution; Mesoudi & Thornton, 2018). However, this is certainly not the status that somebody would ascribe to what is understood as knowledge in individual differences research (Hunt, 2000; Wilhelm & Schroeders, 2019). Roughly speaking, in individual differences research, knowledge is understood as being accumulated over a lifetime (e.g., Cattell, 1987) and as one of many building blocks that constitute intelligence (e.g., Wilhelm & Kyllonen, 2021). This building block is subsumed under the umbrella term of crystallized intelligence (Gc; c.f., Cattell, 1971), which further includes several other language-related abilities (e.g., Carroll, 1993; Schneider & McGrew, 2018).

Research regarding fact knowledge (i.e., declarative knowledge) as an important ingredient of Gc (Wilhelm & Schroeders, 2019) has gained some considerable attention over the course of the last 25 or so years (e.g., Beier & Ackerman, 2003; Hunt, 2000; Rolfhus & Ackerman, 1996, 1999; Rusche & Ziegler, 2022, 2023; Schipolowski et al., 2014; Schroeders, Wilhelm, et al., 2016; Schroeders, Schipolowski, et al., 2016; Schroeders et al., 2021; Steger et al., 2019; Tucker-Drob & Salthouse, 2008; Watrin, Schroeders, et al., 2022; Watrin et al., 2023; Wilhelm & Kyllonen, 2021, etc.). This attention has been necessary because declarative knowledge is relatively understudied compared to other facets of human cognition. Yet, individual differences in Gc, including individual differences in fact knowledge, play an essential role on numerous applied occasions, for example, encompassing contemporary educational assessments like international PISA tests (e.g., OECD, 2013), college admission tests (e.g., Kuncel & Hezlett, 2007; Watrin, Geiger, et al., 2022), and regular school or

university exams. Moreover, such occasions also involve health-related fields where assessments for the identification of neuropsychological disorders are often, among others, based on Gc-related performances that probe semantic memory (c.f., Strauss et al., 2006; Tulving, 1972).

Despite interesting studies regarding declarative knowledge and Gc on both the theoretical and empirical end, there are still open questions that deserve empirical attention. One pragmatic question, for instance, is whether different test modes (e.g., self-reports vs. maximal performance tests) used for measuring declarative knowledge tap the same underlying construct to the same extent, which refers to a classical multi-trait-multi-method problem of psychometrics (Campbell & Fiske, 1959). From this perspective, research on a construct is supposed to transcend the methods of inquiry we apply because the latent construct is supposed to exist independently of the specificity of the measurement instruments that are seeking to capture it (Cronbach & Meehl, 1955). However, not only does it remain an open question whether declarative knowledge can be interpreted as a classic latent construct (e.g., Kan et al., 2011) per se, which would be a presupposition for the above-mentioned multi-trait-multi-method perspective, but also have there been proposed various test modes and measurement instruments for the assessment of declarative knowledge in the literature and it is not warranted that all of these instruments capture what is expected to be measured to the same extent. For example, both self-reports of knowledge (e.g., Rolfhus & Ackerman, 1996) and more conventional tests of maximal performance (e.g., multiple-choice tests; e.g., Schipolowski et al., 2014) have been used to study declarative knowledge. It is unclear whether all proposed test modes maintain the consensual psychometric standard that would be appropriate and expected. Along these test modes, several research questions can be articulated, which are addressed by the dissertation at hand and will be shortly summarized in the following:

First, although self-reports of knowledge might be considered an economic measurement instrument for declarative knowledge, which otherwise has to be measured under the full consideration of various aspects of maximal performance research (c.f.,

Cronbach, 1949), they are known to be susceptible to multiple psychometric shortcomings and biases. For instance, self-reports of knowledge are prone to overclaiming (e.g., Hülür et al., 2011, 2014), that is, the claiming of knowledge one does not or cannot have (e.g., Paulhus et al., 2003). This seriously calls into question the utility of self-report measures of knowledge for diagnostic purposes (c.f., Freund & Kasten, 2012; Zell & Krizan, 2014) and therefore deserves empirical attention. In the same vein, it is an open question of how far such self-report test instruments are to be placed within the existing ability structure, capitalizing on actual tests of maximal performance. To this end, I aimed to study self-reported knowledge and related overclaiming tendencies with regard to their nomological net, including related ability and personality measures.

Second, maximal performance knowledge tests (Cronbach, 1949) are frequently considered to tap different underlying cognitive processes, which might be contingent on the characteristics of the applied test formats (e.g., Becker & Johnston, 1999; Hickson & Reed, 2011; Krieg & Uyar, 2001). For instance, more restricted response formats like multiple-choice tests might only require test takers to recognize the correct response to a given question. In contrast, less restricted response formats like open-ended knowledge questions might involve more sophisticated cognitive processes like proactively retrieving the sought information from long-term memory (e.g., Tulving & Watkins, 1973). If the methods of measurement for assessing declarative knowledge altered the cognitive processes involved, this would have serious psychometric consequences, such as distorted rank orders of individuals across response formats, although the same construct—declarative knowledge—was supposed to be measured in the first place. For that reason, I aimed at psychometrically comparing three competing response formats, including both multiple-choice and instantiations of constructed-response formats that supposedly differ in the cognitive processes they involve. With that, I aimed at shedding light on the idea that individual differences in declarative knowledge might hinge upon the administered response formats.

Third, if we assume that declarative knowledge is an end product of a learning process (c.f., Cattell, 1971, 1987) in the form of a crystallized information structure in long-term

memory (e.g., Cattell, 1971; Hunt, 2000), then we must also assume that there are other kinds of processes or abilities that grant us access to these particular information; otherwise, the stored information "just sits there in memory" (Hunt, 2000, p. 124). Hence, it makes sense to study the nomological net of Gc, and particularly declarative knowledge, in consideration of broad retrieval ability (Gr), which describes the ability of retrieving, restructuring, and (re)producing information from long-term memory (e.g., Schneider & McGrew, 2018). Further, the presumed dependence of Gc performance on related factors such as Gr, which are arguably important for differential levels of Gc to become visible in the first place, question the relative interpretation of Gc as an independent broad ability factor in contemporary models of intelligence (e.g., Kan et al., 2011; McGrew et al., 2023). In contrast to that, the same idea holds for Gr because if no information is stored in long-term memory (i.e., Gc), nothing is there to be retrieved. This poses the interesting question of how previous classifications of Gc and Gr might be continued in the future. Under the assumption that Gr poses a pivotal ingredient of Gc performance and vice versa, the final study of this dissertation thus investigates Gr in a multivariate study, including a selection of several cognitive ability covariates to shed light on the cognitive underpinnings of Gr performance, which in turn should also inform us about the cognitive underpinnings of Gc.

Taken together, in this cumulative dissertation, I present four manuscripts that all, by and large, revolve around the questions of how individual differences in declarative knowledge can and should be measured and how our understanding regarding Gc could be furthered by considering its interdependence from related ability constructs like Gr. In the following sections, I will outline the theoretical background that guided this dissertation by a) embedding the concept of declarative knowledge into contemporary theories of intelligence, by b) describing challenges in the measurement of knowledge, and c) by laying out the cognitive underpinnings that might drive individual differences in declarative knowledge tests.

Closing in on (Declarative) Knowledge

Knowledge is closely intertwined with research efforts set out to understand and explain human cognition. From the perspective of intelligence research, there are mainly two types of knowledge that are distinguished with respect to their substance (Cattell, 1971; for a review of further types of knowledge, see de Jong & Ferguson-Hessler, 1996). Declarative knowledge refers to knowledge about static facts (i.e., fact knowledge); for example, knowing the capital of Sweden or who articulated the categorical imperative. As such, declarative knowledge could also be understood as semantic memory (c.f., Tulving, 1972), that is, as the segment of long-term memory that is conceived as a network of interconnected facts (Collins & Loftus, 1975; Smith et al., 1974). In contrast to that, procedural knowledge refers to knowledge that describes how certain things are done and can often not be sufficiently verbalized; for example, how a flat tire is changed correctly or how children can be smoothed easily. Notably, both types of knowledge exist under the condition that the information units referring to them are stored somewhere in long-term memory after they have been learned. It was stated that "knowledge is exemplified by the recall or recognition of declarative facts and by the demonstration of procedural skills" (Rolfhus & Ackerman, 1999, p. 511). This statement grasps a fundamental idea regarding knowledge: Knowledge—of whatever kind—has to be understood as a store of information units, which were learned and filed at a particular point in time, and which can be reaccessed for further use. This idea of knowledge being a cumulative store is not new (e.g., Cattell, 1943, 1971, 1987; McGrew, 2005) and shares the basal notion that is also characterizing crystallized intelligence (Gc) as a pivotal first-order factor of human cognitive abilities (e.g., Carroll, 1993; Schneider & McGrew, 2018; Wilhelm & Kyllonen, 2021). Thus, in order to understand the concept of knowledge, it is crucial to consider Gc.

In the recent literature, Gc is "typically described as a person's breadth and depth of acquired knowledge of the language, information, and concepts of a specific culture, and / or the application of this knowledge" (McGrew, 2009, p. 5). In addition to that, it was said that "Gc is primarily a store of verbal or language-based declarative (knowing "what") and

procedural (knowing "how") knowledge, acquired through the "investment" of other abilities during formal and informal educational and general life experiences" (McGrew, 2005, p. 151). With that, Gc is contrasted with Gkn (domain-specific knowledge; McGrew, 2005; Rusche & Ziegler, 2023), with the subtle but important difference between both constructs being that "Gc primarily reflects general knowledge via the experience of cultural universials" (McGrew, 2005, p. 152), whereas Gkn is understood as an "individual's breadth and depth of acquired knowledge in specialized (demarcated) domains that typically do not represent the general universal experiences of individuals in a culture" (McGrew, 2005, p. 152). Please note, however, that a clear distinction between both factors is difficult because the assignment of knowledge questions to different more general, or more specific domains is not necessarily straightforward.

The definition of Gc is akin to the so-called pedagogical method of intelligence research proposed almost a century earlier (Binet & Simon, 1904), which capitalized on studying intelligence with regards to finished learning processes (i.e., acquired knowledge, e.g., in school, which should be deemed fairly universal across individuals). In contrast to the pedagogical method, the psychological method (Binet & Simon, 1904) focused on more flexible thinking that might be translated into what is nowadays understood as reasoning ability or fluid intelligence (Gf, e.g., Carroll, 1993; Wilhelm, 2005). Most famously, the distinction between Gc and Gf was articulated in work (e.g., Cattell, 1941, 1943; Hebb, 1941; c.f., Brown, 2016) that developed into what is nowadays known as the investment theory of cognitive abilities (e.g., Cattell, 1963). With that, Cattell joined the distinction between product and process aspects of intelligence. This distinction also found appeal in extensions of the theory (e.g., Intelligence-as-Process, Personality, Interests, and Intelligence-as-Knowledge, Ackerman, 2000, 1996; extended Gf-Gc theory, Horn & Noll, 1997; Openness-Fluid-Crystallized-Intelligence Model, Ziegler et al., 2012) that all emphasize the role of conative factors for the development of the product aspects of intelligence. Also, related areas of research built on this (e.g., in developmental psychology, where it is distinguished between pragmatic and mechanic aspects of cognition, respectively; c.f., Baltes et al., 1999). This is

important because the distinction of Gc and Gf through the lens of the investment theory informs us about the concept of Gc sensu Cattell (1971) and will shed light on how declarative knowledge must inevitably be understood as a pivotal part of Gc (e.g., Wilhelm & Kyllonen, 2021; Wilhelm & Schroeders, 2019).

The main idea of the investment theory is that throughout an individual's life, fluid abilities, such as Gf, are invested across a multitude of learning opportunities and that the end product of these learning processes is mirrored in what is understood as crystallized intelligence. Therefore, Gc is often deemed the second most important factor of human cognitive abilities after Gf (c.f., Wilhelm & Schroeders, 2019). In fact, Gf was found to influence Gc across the life span (Beier & Ackerman, 2001), and whereas Gf reaches its peak in early adulthood followed by a steady decline, Gc continuously rises and does not decline until late adulthood (Ackerman, 1996, 2000; Cattell, 1971; Horn & Donaldson, 1980; Horn & Hofer, 1992). Thus, "crystallized" intelligence indicates that all performances that are subsumed under this term are "fixated and non-flexible" (Schipolowski, 2014, p. I-8) in the sense that they depict an end-product, that is, some crystallized information structures that are the result of learning, steadily ongoing acculturation processes (e.g., Ackerman, 2000), and of intellectual investment traits (e.g., von Stumm, 2013; Wilhelm et al., 2003; Ziegler et al., 2012) including interests (e.g., Su et al., 2009; Tobias, 1994). In other words, the "term crystallized is meant to imply freezing in a specific shape of what was once fluid" (Cattell, 1971, p. 118). This idea relates well to what is understood as declarative knowledge; namely static facts that are, once learned, readily available for retrieval from long-term memory unless they are forgotten (Cattell, 1971). In fact, for Cattell, who originally introduced the term "crystallized intelligence" (Cattell, 1943), Gc was the entirety of knowledge that humans encounter and accumulate during their lifetime (Cattell, 1971). From this perspective, it becomes apparent that a key component of Gc should be understood as declarative knowledge regarding broad and general content domains (e.g., knowledge about humanities), which can be learned in both curricular and extracurricular contexts (Rolfhus & Ackerman, 1996; Schipolowski et al., 2014; Steger et al., 2019; Wilhelm & Kyllonen, 2021).

However, this perspective stands in stark contrast to various other interpretations regarding Gc that have omitted or neglected declarative fact knowledge as an essential part of Gc. This is aptly revealed by reviewing past measurement efforts regarding Gc. Although a broad Gc factor is proposed in all extensions of earlier intelligence theories (e.g., the Three-Stratum-Theory, Carroll, 1993; the Cattel-Horn-Carroll-Theory, McGrew, 2005, 2009; Schneider & McGrew, 2018), conventional measurement instruments that are widely in use do only seldom include assessments of fact or declarative knowledge and focus instead on tests of verbal or language-related abilities, such as vocabulary tests (e.g., Schipolowski et al., 2014) or tests of reading comprehension (e.g., Hartung et al., 2017; Thorsen et al., 2014). The latter might be, in turn, confounded with Gf requirements (Hartung et al., 2017; Wilhelm & Schroeders, 2019) – something that should be avoided to maximize the distinction between Gc and Gf (Carroll, 1993), which is desired in order to study relatively pure factors of cognitive abilities. Furthermore, relying on verbal ability tests for measuring Gc was deemed inadequate because these abilities would not cover the broadness of the Gc factor as it was initially conceptualized (Cattell, 1971; Wilhelm & Kyllonen, 2021).

Taken together, assessing declarative knowledge can help to capture Gc as it was initially proposed – as a broad factor, as distinct as possible from Gf, and shaped by lifelong learning and acculturation processes. In fact, declarative knowledge tests were found to supplement tests of verbal and language-related abilities (e.g., Horn, 1965; Horn & Cattell, 1966), which is also apparent by a robust latent correlation (r = .91) between verbal abilities and declarative knowledge (Schipolowski et al., 2014). These empirical findings propel the argument that declarative knowledge tests can serve as a good marker for Gc, additionally to existing measurement instruments focusing on more language-related abilities (Amthauer et al., 2001; Beauducel & Kersting, 2002). In this context, it is notable to mention that under the consideration of various content domains (Ackerman, 1996), which reflect conceptually broad assessments of declarative knowledge (Ackerman, 2000), a higher-order factor of Gc covering several broad knowledge domains modeled as first-order factors (e.g., natural sciences, life sciences, humanities, and social sciences; likely depending on the types of indicators that are

chosen) that individuals arguably accumulate over their life time can be identified (c.f., Steger et al., 2019). Thus, earlier calls to measure Gc as broad as possible (e.g., Cattell, 1971) find empirical support that is regularly replicated (e.g., Goecke, Staab, et al., 2022; Schroeders et al., 2021; Watrin, Schroeders, et al., 2022).

Measurement of Declarative Knowledge

Not without reason has declarative knowledge been described as the "dark matter of adult intelligence" research (Ackerman, 2000, p. 69), or Gc, in a broader sense, been scolded a "wallflower" (Hunt, 2000, p. 124). Studying individual differences in knowledge seems straightforward, namely simply assessing how individuals differ in the knowledge they are able to display through probing it with suitable measurement instruments (Cronbach, 1975). However, the essential concept of the investment theory (Cattell, 1987) and related extensions (e.g., Ackerman, 1996, 2000; Ziegler et al., 2012) illustrate clearly that the accumulation of knowledge is an inherently idiosyncratic process that results in numerous challenges regarding a clean and pure assessment of declarative knowledge (Wilhelm & Schroeders, 2019). For example, it was argued that the assessment of declarative knowledge becomes more complex as learning opportunities diversify, especially after formal education (Cattell, 1971).

Declarative knowledge tests that are set out to capture the broadness of the underlying learning opportunities and therefore paint a holistic picture of an individual's knowledge are supposed to cover the entirety of knowledge there is, but this assumption is ill-fated, as it is an impossible task to design a measurement instrument that lives up to this expectation (Ackerman, 2000; Steger et al., 2022). Thus, belonging to challenges discussed and investigated empirically time and again are, for example issues of item sampling (e.g., short vs. long tests, Schipolowski et al., 2013; domain-specific vs. domain-general tests, Achaa-Amankwaa et al., 2022; Ackerman, 2000; Ackerman & Rolfhus, 1999; Rolfhus & Ackerman, 1996; Rusche & Ziegler, 2023; Steger et al., 2019; Steger et al., 2022; Watrin et al., 2022; c.f., Wilhelm & Kyllonen, 2021), person sampling (e.g., young vs. old age, Schroeders et al., 2016; Schroeders et al., 2021; sex differences, Schroeders et al., 2015; country-differences, Watrin et

al., 2023), and test contexts (e.g., proctored vs. unproctored tests, Steger et al., 2021), but also important moderating effects of learning opportunities throughout one's life and the access to the same, which might reflect critical interdependencies between investment traits and knowledge acquisition (e.g., Tucker-Drob & Briley, 2014). Although these challenges must certainly be acknowledged as such, and the research efforts empirically investigating them with rigor deserve applause, a more extensive literature review regarding this specific challenge is beyond the scope of this dissertation.

Instead, it is also important to consider more basal questions that pertain to measurement-related issues and should be considered just as pressing. One of these questions concerns the basic distinction between maximal ability and typical behavior (e.g., Cronbach, 1949) in the measurement of declarative knowledge. Whereas tests of maximal ability "seek[s] a measure of the best performance the person can produce" (Cronbach, 1949, p. 13), tests of typical behavior do not probe test takers' maximal cognitive effort but instead focus on tests where "'good' or 'best' performance is inappropriate" (Cronbach, 1949, p. 14): for example, self-reports in personality research. The distinction between both types of behavior also finds reflection in the data types provided in Raymond Cattells' "Three Media for Personality Measurement" (Cattell, 1958, p. 286). While self-reports, which are usually thought to assess typical behavior, are counted to the data type Q-data (i.e., questionnaire data) or L-Data (i.e., life data), tests of maximal performance are assigned to T-data (i.e., test data).

Because declarative knowledge is understood as an important ingredient of Gc, and Gc, in turn, is understood to be a pivotal factor of general intelligence, declarative knowledge should ostensibly be measured by tests of maximal performance (i.e., T-data). More than that: If we do not measure declarative knowledge with tests of maximal performance, then what do we measure in the first place? Tests of maximal performance provide inherent advantages over the use of self-report measures, and displaying knowledge must be considered an ability. Self-report measures are frequently criticized for their psychometric shortcomings, which include that they are fakeable (c.f., Geiger et al., 2018), they presume an introspection ability of the individual test taker that is often ungrounded (c.f., Baas et al., 2015), and they lack the

possibility of being evaluated for their veridicality, which is considered a vital standard in cognitive ability testing (c.f., Wilhelm & Schroeders, 2019). Lastly, self-reports are generally known to be biased by a variety of distortions (e.g., Howard, 1994; Schwarz, 1999).

Nevertheless, self-reports are economic in application, and hence the idea of approximating a construct of maximal performance, like declarative knowledge through the use of self-reports, seems appealing. In fact, there is a large body of literature capitalizing on this exact idea (for an overview, c.f. Freund & Kasten, 2012). For example, self-reports are used to assess how individuals would rate their own abilities (e.g., general intelligence or specific facets like verbal abilities), either in absolute terms or relative to others. Although items like "How smart do you think you are?" do then not probe actual typical behavior, measurement instruments deploying such items should still be regarded as belonging to Q-data and might then be called self-reports of (cognitive) ability or self-estimates of (cognitive) ability. Although it might seem appealing to use such measurement instruments to study variables that are actually considered an ability, the empirical picture regarding such self-reports is disillusioning and plagued by obvious distortions (e.g., Ackerman & Wolman, 2007). Individuals are generally found to be poor in giving accurate self-estimates of their abilities; correlations between perceived and actual abilities do, on average, usually not exceed $r = .32$ (c.f., Freund & Kasten, 2012, and Zell & Krizan, 2014 for recent meta-analytical evidence regarding this claim; but see also Mabe & West, 1984). This also indicates a lack of convergent validity of self-reports of ability (e.g., Neubauer & Hofer, 2021) and thus calls into question the use of such measurement instruments, which has also resonated in related literature (e.g., literature regarding the Dunning-Kruger effect, which has been viewed critically in recent years; Gignac & Zajenkowski, 2020; Hofer et al., 2022; Kruger & Dunning, 1999).

In contrast to these considerations, self-reports also find use in the study of (domain-specific) declarative knowledge (e.g., Cotter et al., 2023; Raubenheimer, 1925; Rolfhus & Ackerman, 1996; Silvia & Berg, 2011), and were initially even used to inform critical research regarding the dimensionality of declarative knowledge (Rolfhus & Ackerman, 1996). In these

cases, test takers are usually not asked to estimate their level of knowledge but to indicate their familiarity with certain words, terms, or concepts that might be assigned to a specific knowledge domain (e.g., chemistry). As such, self-reports of knowledge can also be considered to tap what a person knows, but as opposed to tests of maximal performance, there is no control in place that evaluates whether the statement of a test taker can be trusted or not. For example, in a study by Rolfhus & Ackerman (1996), participants were asked to rate on 552 items "how much information they could produce on the subject, relative to other people they knew" (Rolfhus & Ackerman, 1996, p. 177) on a scale from "*not at all knowledgeable*" to "*extremely knowledgeable*". Using this design, the authors identified six broad knowledge domains and hypothesized that "objective ability measures would correlate with the appropriate self-report knowledge domains on the basis of content similarity" (Rolfhus & Ackerman, 1996, p. 176). However, they did not include such a domain-specific ability measure. A more recent example deployed a self-reported knowledge scale regarding art knowledge (Cotter et al., 2023), where participants were probed with 36 single art terms or persons and asked to rate their knowledge on a scale from "*I have never heard of this artist or term*" to "*I can talk intelligently about this artist or idea in art*" (Cotter et al., 2023, p. 2). The authors discussed that "how much people know about art is fundamental to how they experience and encounter it" (Cotter et al., 2023, p. 16) and concluded that their versions of the "Aesthetic Fluency Scale" "are promising tools for art researchers" (Cotter et al., 2023, p. 17). This conclusion should be scrutinized critically, however. Because self-reports of ability are known to be inaccurate, it is also reasonable to assume that self-reported knowledge is subject to unwanted distortions.

Evidence for this assumption comes from another body of research, which was originally concerned with examining and controlling for self-enhancement tendencies in self-reports (Paulhus et al., 2003). In this line of research, the concept of overclaiming (i.e., knowledge purporting), that is, claiming more knowledge than one has or actually can have, was re-introduced (c.f., Raubenheimer, 1925 who studied knowledge purporting in order to identify delinquent children; Phillips & Clancy, 1972 who studied knowledge purporting in

the vein of consumer research). The basic idea of overclaiming is that test takers might exaggerate their knowledge on self-reports of knowledge by claiming familiarity with terms they do not actually know—or cannot know, because they do not actually exist. As such, overclaiming is usually measured with so-called Overclaiming Questionnaires (OCQ; Paulhus et al., 2003), which employs two types of items: *reals* (i.e., words, concepts or terms that actually exist; e.g., *psychology*) and *foils* (i.e., words, concepts or terms that are entirely made up; e.g., *epocriticism*).

Akin to the above-introduced self-reports of knowledge, test takers of OCQs are usually asked to rate their familiarity with the provided terms on a scale from "*Never heard of it before"* to "*I could give an exact description*" (c.f., Goecke et al., 2020). Test takers stating familiarity with foils can be considered overclaimers (Bing et al., 2011; Fell et al., 2017; Goecke et al., 2020; Goecke, Schroeders, et al., 2022; Hülür et al., 2011, 2014), whereas rating familiarity with reals might be considered as self-reported knowledge or tapping semantic memory (e.g., Tulving, 1972). As such, OCQs or self-reports of knowledge based on single words and familiarity ratings are very much akin. There is a large body of literature regarding overclaiming and OCQs, and one of the ever-occurring main results of this literature seems to be the finding that self-reported knowledge is severely inflated by overclaiming tendencies (which might occur due to various reasons; see Goecke et al., 2020 in this dissertation for an overview).

Taken together, it is questionable how far self-reports of knowledge are to be placed within the conventional ability structure from a mono-trait-multi-method perspective (c.f., Campbell & Fiske, 1959), but as self-reports are used to assess knowledge from time to time, it seems important to test their capability of doing so. In addition to that, it is equally important to consider more conventional maximal performance measures of declarative knowledge and Gc and start investigating the cognitive processes involved in displaying knowledge tests tapping these constructs (c.f., Frischkorn et al., 2022).

Declarative Knowledge as a Vault of Information

For shedding light on the cognitive underpinnings of declarative knowledge, it is important to consider the measurement instruments normally used for assessing declarative knowledge. Declarative knowledge has a longstanding tradition of being assessed with multiple-choice formats (e.g., Ackerman, 1996). Competing response formats, like constructed-response formats (e.g., Traub, 1993), are less frequently used but are time and again hypothesized to tap different underlying cognitive processes as compared to multiple-choice formats when responding to fact knowledge questions (for an overview, see Goecke et al., 2022 in this dissertation). For example, whereas multiple-choice formats are often viewed as measurement instruments tapping pure recognition because several response alternatives are already provided for test takers, constructed-response formats might put stronger demand on actual recall because the correct response to a knowledge question has to be identified from a manifold of existing information in long-term memory and then proactively used (e.g., Tulving & Watkins, 1973). Hence, understanding whether declarative knowledge might change as a function of the applied response format seems informative regarding the cognitive processes that might be responsible for individual differences in declarative knowledge. With that, fundamental questions of cognitive psychology (i.e., discontinuity hypothesis, Tulving & Watkins, 1973) can be transferred to the study of knowledge. In line with that, it is worth considering the nature of crystallized intelligence again.

What else really is declarative knowledge than a crystallized information structure that sits somewhere in memory and awaits carving? Knowledge might be a "dark matter" (Ackerman, 2000), but from a more cognitive perspective, knowledge might also be described as the brown coal of the human mind. The similarities imposed by this metaphor are startling. Brown coal has been accumulated over the earth's history and is the preliminary end product of a long process of plant carbonization over the course of earth ages. This end product now lies hidden in the earth's mantle, waiting to be extracted by human beings and used for energy production until viable alternatives are available. Without humans, however, it would just sit there (and possibly become more over time). Now consider again the perspective that

declarative knowledge corresponds to a purely crystallized end product that is created by learning (c.f., investment theory, Cattell, 1987) – in whatever form. From this perspective, knowledge "just sits there in [long-term] memory" (Hunt, 2000, p. 124) after the actual learning process (e.g., through studying in school, or through talking to strangers in another country, or through observation, etc.) is completed; the learned information is then stored in memory (e.g., Cohen et al., 1985), but not yet used further, for example by applying it in an exam. Now, through different learning opportunities, different life choices, distinct coincidences during our time on this planet, or different interpretations of our realities, knowledge and its corresponding structure might look a little different in each individual's memory (Siew & Guru, 2022). However, if the assumption that knowledge, in general, is stored in some kind of crystallized information structure is at least partly correct, all individuals would be in need of some tool that grants them access to this vault of information for further use; just like brown coal requires humans with heavy equipment to serve its questionable purpose. Some of the information sitting in memory might be carved for specific purposes, but some information might never be touched upon again after it has been filed.

Due to that, it is important to consider the cognitive underpinnings of knowledge, or in other words, those constructs and variables that help individuals gain access to the information they have stored in memory. Hence, bringing together nomological network theory (Cronbach & Meehl, 1955) and recent calls for investigating cognitive processes involved in individual differences in intelligence (Frischkorn et al., 2022) is warranted. Whereas the first describes that latent entities such as Gc might be viewed as constructs that can be empirically delineated through their relationship with other constructs through psychometric analyses (c.f., convergent and discriminant validity; Campbell & Fiske, 1959), the latter urges intelligence researchers to study intelligence process-oriented (Frischkorn et al., 2022).

One candidate construct that seems helpful for understanding how individual differences in maximal performance tests tapping declarative knowledge come about is broad retrieval ability (Gr in the CHC model of intelligence; e.g., Schneider & McGrew, 2018). Gr

can be defined as "the rate and fluency with which individuals can selectively retrieve and produce information stored in long-term memory" (i.e., knowledge) (Schneider & McGrew, 2018, p. 102). This definition underlines not only the presumed importance of Gr for differential levels of declarative knowledge but also the reciprocal relationship of both constructs because it refers to what was above described as a "tool" that grants individuals access to their stored knowledge. As such, Gr and its underlying cognitive processes might be understood as the mechanical part (Wilhelm & Schroeders, 2019) that complements declarative knowledge specifically, or Gc more generally (e.g., McGrew et al., 2023; Schneider & McGrew, 2018). In turn, Gc or declarative knowledge might be understood as the "knowledge-driven aspect" (Wilhelm & Schroeders, 2019, p. 261) that fuels Gr because Gc is the pile of knowledge lying around, whereas Gr is the process of retrieving knowledge. At last, if one of both constructs were missing in the equation, possibly no information from long-term memory could be retrieved whatsoever. Hence, to comprehend Gc to an exhaustive extent requires us also to study Gr.

Previous differential research on Gr is usually based on simple count variables (i.e., fluency; c.f., Unsworth, 2019). However, various more sophisticated models of performance in semantic fluency tasks distinguish between several cognitive processes that supposedly underly semantic retrieval performances (c.f., Rosen & Engle, 1997; Unsworth et al., 2011) and might thus also be important for individual differences in declarative knowledge, as responding to knowledge questions demands test takers to retrieve the required response from long-term memory – a requirement that is shared by both typical indicators of declarative knowledge and typical indicators of semantic retrieval ability. For example, a previously suggested four-component model of semantic retrieval ability (Rosen & Engle, 1997) includes four components that could be understood as the main ingredients of the process of retrieving selected information from long-term memory: components of spreading activation, monitoring, suppression, and self-generation of category cues. These components illustrate the dynamic interplay between both strategic and automatic search components (Unsworth et al., 2011). Importantly, the components of the model by Rosen & Engle (1997)

can be mapped onto cognitive constructs pertaining to factors of the CHC model that, in turn, could be understood as satisfactory empirical instantiations of theory-driven determinants of differential Gr levels; and thus of differential levels in declarative knowledge (see Goecke et al., 2023 in this dissertation).

Overview of the Dissertation Manuscripts

Understanding any psychological entity is closely intertwined both with the question of measurement (Campbell & Fiske, 1959; Cronbach, 1949) and with the idea of a nomological net surrounding the construct in question (e.g., Cronbach & Meehl, 1955). In line of this dissertation, I thus present four manuscripts that revolve around the investigation of individual differences in declarative knowledge by focusing on the question of how we can and should measure and interpret declarative knowledge. The four presented manuscripts are related through their immanent focus on studying variables that are connected to declarative knowledge from various perspectives, namely test modes and the nomological net of knowledge. In the next sections, I will shortly summarize each of the manuscripts before they are presented as the main product of this dissertation. Inherently, the presented manuscripts will all serve to address specific research questions, which I will integrate into a broader framework of the existing literature in the epilogue of this dissertation.

Taken together, the following three overarching questions are tackled by this dissertation:

1. How is self-reported knowledge to be placed within the existing ability structure?
2. How do different response formats affect the measurement of declarative knowledge?
3. Which cognitive abilities explain individual differences in broad retrieval ability?

Manuscript I: Testing Competing Claims About Overclaiming

In the first manuscript of this dissertation (Goecke et al., 2020), I studied self-reported knowledge and overclaiming as measured by OCQs (Paulhus et al., 2003) in a multivariate study with adults. In this context, self-reported knowledge is understood as a potential proxy of actual knowledge, and overclaiming is seen as a potential threat to the validity of self-reports of knowledge. The manuscript provides competing theoretical views on what

overclaiming behavior actually represents. On the basis of these perspectives, I derive hypotheses about the place of overclaiming and self-reported knowledge in their respective nomological net and test the perspectives by means of confirmatory factor analysis. Thereby the study includes constructs from both intelligence and personality research, which is useful for testing the competing perspectives.

Manuscript II: The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children

In the second manuscript (Goecke et al., 2022), I built upon the ideas of the first manuscript and extended the research questions presented in the first manuscript regarding the nomological net of overclaiming and self-reported knowledge to a sample of young children. As in manuscript I, I test several competing perspectives regarding overclaiming and self-reported knowledge and provide the results of a multivariate study, including various measures of intelligence and personality, which are derived by means of confirmatory factor analysis. The study adds to the sparse literature on overclaiming and self-reported knowledge in children.

Manuscript III: Stop Worrying About Multiple-Choice: Knowledge Does Not Change With Response Format

In the third manuscript (Goecke et al., 2022), I investigated the potential psychometric consequences of applying three different response formats (multiple-choice vs. open-ended vs. cued open-ended) for the assessment of declarative knowledge. Within the manuscript, two large online studies are presented, which provide data to put the potential psychometric consequences of applying different response formats to a critical test. Further, the study provides a discussion of potential cognitive underpinnings of responding to knowledge questions and helps to mitigate frequently brought forward reservations regarding multiple-choice response formats.

Manuscript IV: Driving Factors of Individual Differences in Broad Retrieval Ability:* Gr *Is More Than the Sum of Its Parts

In the last manuscript of this dissertation (Goecke et al., 2023), I studied broad retrieval ability (Gr) in a fourth multivariate study and tested whether Gr is best understood as a mere linear function of other related ability constructs, including mental speed, working memory capacity, secondary memory, and crystallized intelligence, or whether Gr should be interpreted as more than the sum of its parts. To this end, I summarize explanatory models for performance in Gr provided in the previous literature and derive testable hypotheses, which can be studied by means of structural equation models. Further, I tested whether specific retrieval affordances, such as short vs. long periods of retrieval periods could be psychometrically distinguished from the more general retrieval affordances.

On the following pages, the aforementioned manuscripts will be presented in full before their major findings and the implications thereof will be summarized and brought into a broader perspective, including current theories and outlines for future work in the epilogue of this dissertation.

References

Achaa-Amankwaa, P., Steger, D., Wilhelm, O., & Schroeders, U. (2022). *Public Events Knowledge in an Age-Heterogeneous Sample: Reminiscence Bump or Bummer?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/yg3mf

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence*, *22*(2), 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*(2), 69–84. https://doi.org/10.1093/geronb/55.2.P69

Ackerman, P. L., & Wolman, S. D. (2007). Determinants and validity of self-estimates of abilities and self-concept measures. *Journal of Experimental Psychology: Applied*, *13*(2), 57–78. https://doi.org/10.1037/1076-898X.13.2.57

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

Baas, M., Koch, S., Nijstad, B. A., & De Dreu, C. K. W. (2015). Conceiving creativity: The nature and consequences of laypeople's beliefs about the realization of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *9*(3), 340–354. https://doi.org/10.1037/a0039420

Baltes, P. B., Staudinger, U. M., & Lindenberger, U. (1999). Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology*, *50*(1), 471–507. https://doi.org/10.1146/annurev.psych.50.1.471

Beauducel, A., & Kersting, M. (2002). Fluid and Crystallized Intelligence and the Berlin Model of Intelligence Structure (BIS). *European Journal of Psychological Assessment*, *18*(2), 97–112. https://doi.org/10.1027//1015-5759.18.2.97

Becker, W. E., & Johnston, C. (1999). The Relationship between Multiple Choice and Essay Response Questions in Assessing Economics Understanding. *Economic Record*, *75*(4), 348–357. https://doi.org/10.1111/j.1475-4932.1999.tb02571.x

Beier, M. E., & Ackerman, P. L. (2001). Current-events knowledge in adults: An investigation of age, intelligence, and nonability determinants. *Psychology and Aging*, *16*(4), 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*(2), 439–448. https://doi.org/10.1037/0022-3514.84.2.439

Binet, A., & Simon, T. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'année Psychologique*, *11*, 191–244. https://doi.org/doi:10.3406/psy.1904.3675

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

Brown, R. E. (2016). Hebb and Cattell: The Genesis of the Theory of Fluid and Crystallized Intelligence. *Frontiers in Human Neuroscience*, *10*, 1–10. https://doi.org/10.3389/fnhum.2016.00606

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, *38*(7). https://doi.org/10.1037/h0059973

Cattell, R. B. (1943). The Measurement of Adult Intelligence. *Psychological Bulletin*, *40*(3), 153–193. https://doi.org/10.1037/h0059973

Cattell, R. B. (1958). What is "objective" in "objective personality tests?" *Journal of Counseling Psychology*, *5*(4), 285–289. https://doi.org/10.1037/h0046268

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22. https://doi.org/10.1037/h0046743

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Houghton Mifflin Company.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. Elsevier Science Publishers B.V.

Cohen, N. J., Eichenbaum, H., Deacedo, B. S., & Corkin, S. (1985). Different Memory Systems Underlying Acquisition of Procedural and Declarative Knowledge. *Annals of the New*

York Academy of Sciences, *444*(1), 54–71. https://doi.org/10.1111/j.1749-6632.1985.tb37579.x

Collins, A. M., & Loftus, E. F. (1975). A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Cotter, K. N., Rodriguez-Boerwinkle, R. M., Christensen, A. P., Fekete, A., Smith, J. K., Smith, L. F., Tinio, P. P. L., & Silvia, P. J. (2023). Updating the Aesthetic Fluency Scale: Revised long and short forms for research in the psychology of the arts. *PLOS ONE*, *18*(2), e0281547. https://doi.org/10.1371/journal.pone.0281547

Cronbach, L. J. (1949). *Essentials of Psychological Testing*. Harper & Brothers, Publishers.

Cronbach, L. J. (1975). Five Decades of Public Controversy Over Mental Testing. *American Psychologist*, *30*(1), 1–14. https://doi.org/10.1037/0003-066X.30.1.1

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

de Jong, T., & Ferguson-Hessler, M. G. M. (1996). Types and Qualities of Knowledge. *Educational Psychologist*, *31*(2), 105–113. https://doi.org/10.1207/s15326985ep3102_2

Fell, C. B., König, C. J., Jung, S., Sorg, D., & Ziegler, M. (2017). Are country level prevalences of rule violations associated with knowledge overclaiming among students? *International Journal of Psychology*, *54*(1), 1–6. https://doi.org/10.1002/ijop.12441

Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, *138*(2), 296–321. https://doi.org/10.1037/a0026556

Frischkorn, G. T., Wilhelm, O., & Oberauer, K. (2022). Process-oriented intelligence research: A review from the cognitive perspective. *Intelligence*, *94*, 1–14. https://doi.org/10.1016/j.intell.2022.101681

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

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence*, *80*, 1–9. https://doi.org/10.1016/j.intell.2020.101449

Goecke, B., Schroeders, U., Zettler, I., Schipolowski, S., Golle, J., & Wilhelm, O. (2022). The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children. *Journal of Personality Assessment*, 1–12. https://doi.org/10.1080/00223891.2022.2144332

Goecke, B., Staab, M., Schittenhelm, C., & Wilhelm, O. (2022). Stop Worrying about Multiple-Choice: Fact Knowledge Does Not Change with Response Format. *Journal of Intelligence*, *10*(4), 1–22. https://doi.org/10.3390/jintelligence10040102

Goecke, B., Weiss, S., Steger, D., Schroeders, U., & Wilhelm, O. (2020). Testing competing claims about overclaiming. *Intelligence*, *81*, 1–11. https://doi.org/10.1016/j.intell.2020.101470

Goecke, B., Weiss, S., & Wilhelm, O. (2023). *Driving Factors of Individual Differences in Broad Retrieval Ability: Gr Is More Than the Sum of Its Parts* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/zy7mp

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

Hebb, D. O. (1941). Clinical evidence concerning the nature of normal adult test performance. *Psychological Bulletin*, *38*(7).

Hickson, S., & Reed, B. (2011). More evidence on the use of constructed-response questions in principles of economics classes. *International Review of Economics Education*, *10*(2), 28–49. https://doi.org/10.1016/S1477-3880(15)30030-X

Hofer, G., Mraulak, V., Grinschgl, S., & Neubauer, A. C. (2022). Less-Intelligent and Unaware? Accuracy and Dunning–Kruger Effects for Self-Estimates of Different Aspects of Intelligence. *Journal of Intelligence*, *10*(1), 1–18. https://doi.org/10.3390/jintelligence10010010

Horn, J. L. (1965). *Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities.* University of Illinois.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253–270. https://doi.org/10.1037/h0023816

Horn, J. L., & Donaldson, G. (1980). Cognitive Development in Adulthood. In O. G. Jr. Brim & J. Kagan (Eds.), *Constancy and Change in Human Development* (pp. 445–529). Harvard University Press.

Horn, J. L., & Hofer, S. M. (1992). Major Abilities and Development in the Adult Period. In R. J. Sternberg & C. Berg (Eds.), *Intellectual Development* (pp. 44–99). Cambridge University Press.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 53–91). The Guilford Press.

Howard, G. S. (1994). Why do people say nasty things about self-reports? *Journal of Organizational Behavior*, *15*(5), 399–404. https://doi.org/10.1002/job.4030150505

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

Hülür, G., Wilhelm, O., & Schipolowski, S. (2014). Corrigendum to "Prediction of self-reported knowledge with over-claiming, fluid and crystallized intelligence and typical intellectual engagement" [Learning and Individual Differences 21 (6) (2011)

742–746]. *Learning and Individual Differences*, *35*, 153–154. https://doi.org/10.1016/j.lindif.2014.07.014

Hunt, E. (2000). Let's hear it for crystallized intelligence. *Learning and Individual Differences*, *12*(1), 123–129. https://doi.org/10.1016/S1041-6080(00)00033-9

Kan, K.-J., Kievit, R. A., Dolan, C., & der Maas, H. van. (2011). On the interpretation of the CHC factor Gc. *Intelligence*, *39*(5), 292–302. https://doi.org/10.1016/j.intell.2011.05.003

Krieg, R. G., & Uyar, B. (2001). Student performance in business and economics statistics: Does exam structure matter? *Journal of Economics and Finance*, *25*(2), 229–241.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized Tests Predict Graduate Students' Success. *Science*, *315*(5815), 1080–1081. https://doi.org/10.1126/science.1136618

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of abilities: A review and meta-analysis. *Journal of Applied Psychology*, *67*(3), 280–296. https://doi.org/10.1037/0021-9010.67.3.280

McGrew, K. S. (2005). The Cattell-Horn-Carroll Theory of Cognitive Abilities: Past, Present, and Future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 136–181). The Guilford Press.

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), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

McGrew, K. S., Schneider, W. J., Decker, S. L., & Bulut, O. (2023). A Psychometric Network Analysis of CHC Intelligence Measures: Implications for Research, Theory, and Interpretation of Broad CHC Scores "Beyond g." *Journal of Intelligence*, *11*(1), 19. https://doi.org/10.3390/jintelligence11010019

Mesoudi, A., & Thornton, A. (2018). What is cumulative cultural evolution? *Proceedings of the Royal Society B: Biological Sciences*, *285*(1880), 20180712. https://doi.org/10.1098/rspb.2018.0712

Neubauer, A. C., & Hofer, G. (2021). Self-estimates of abilities are a better reflection of individuals' personality traits than of their abilities and are also strong predictors of professional interests. *Personality and Individual Differences*, *169*, 1–15. https://doi.org/10.1016/j.paid.2020.109850

OECD. (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD. https://doi.org/10.1787/9789264190511-en

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

Phillips, D. L., & Clancy, K. J. (1972). Some effects of "social desirability" in survey studies. *American Journal of Sociology*, *77*(5), 921–940. https://doi.org/10.1086/225231

Raubenheimer, A. S. (1925). An experimental study of some behavior traits of the potentially delinquent boy. *Psychological Monographs*, *34*, 1–107.

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. https://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*(3), 511–526. https://doi.org/10.1037/0022-0663.91.3.511

Rosen, V. M., & Engle, R. W. (1997). The Role of Working Memory Capacity in Retrieval. *Journal of Experimental Psychology: General*, *126*(3), 211–227. https://doi.org/10.1037/0096-3445.126.3.211

Rusche, M. M., & Ziegler, M. (2022). The interplay between domain-specific knowledge and selected investment traits across the life span. *Intelligence*, *92*, 101647. https://doi.org/10.1016/j.intell.2022.101647

Rusche, M. M., & Ziegler, M. (2023). Measuring Domain-Specific Knowledge: From Bach to Fibonacci. *Journal of Intelligence*, *11*(3), 47. https://doi.org/10.3390/jintelligence11030047

Schipolowski, S. (2014). *Zur Struktur, Messung und Entwicklung kristalliner Intelligenz [On the Structure, Measurement, and Development of Crystallized Intelligence]*. Humboldt-Universität zu Berlin.

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). 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

Schipolowski, S., Wilhelm, O., Schroeders, U., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2013). BEFKI GC-K: A Short Scale for the Measurement of Crystallized Intelligence. *methods, data, analyses*, *7*(2). https://doi.org/10.12758/MDA.2013.010

Schneider, W. J., & McGrew, K. S. (2018). The Cattell–Horn–Carroll Theory of Cognitive Abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 73–163). The Guilford Press.

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., Schipolowski, S., Zettler, I., Golle, J., & Wilhelm, O. (2016). Do the smart get smarter? Development of fluid and crystallized intelligence in 3rd grade. *Intelligence*, *59*, 84–95. https://doi.org/10.1016/j.intell.2016.08.003

Schroeders, U., Watrin, L., & Wilhelm, O. (2021). Age-related nuances in knowledge assessment. *Intelligence*, *85*, 101526. https://doi.org/10.1016/j.intell.2021.101526

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

Schwarz, N. (1999). How the Questions Shape the Answers. *American Psychologist*, *54*(2), 93–105.

Siew, C. S. Q., & Guru, A. (2022). Investigating the network structure of domain-specific knowledge using the semantic fluency task. *Memory & Cognition*. https://doi.org/10.3758/s13421-022-01314-1

Silvia, P. J., & Berg, C. (2011). Finding Movies Interesting: How Appraisals and Expertise Influence the Aesthetic Experience of Film. *Empirical Studies of the Arts*, *29*(1), 73–88. https://doi.org/10.2190/EM.29.1.e

Smith, E. E., Rips, L. J., & Shoben, E. J. (1974). Semantic Memory and Psychological Semantics. *Psychology of Learning and Motivation*, *8*, 1–45. https://doi.org/10.1016/S0079-7421(08)60451-X

Steger, D., Jankowsky, K., Schroeders, U., & Wilhelm, O. (2022). The Road to Hell Is Paved With Good Intentions: How Common Practices in Scale Construction Hurt Validity. *Assessment*, *00*(0), 1–14. https://doi.org/10.1177/10731911221124846

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

Steger, D., Schroeders, U., & Wilhelm, O. (2021). Caught in the act: Predicting cheating in unproctored knowledge assessment. *Assessment*, *28*(3), 1004–1017.

Steup, M., & Neta, R. (2005). *Epistemology* (E. N. Zalta, Ed.; Fall 2020 Edition). https://plato.stanford.edu/archives/fall2020/entries/epistemology/

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary* (Third Edition). Oxford University Press.

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, *135*(6), 859–884. https://doi.org/10.1037/a0017364

Thorsen, C., Gustafsson, J.-E., & Cliffordson, C. (2014). The influence of fluid and crystallized intelligence on the development of knowledge and skills. *British Journal of Educational Psychology*, *84*(4), 556–570. https://doi.org/10.1111/bjep.12041

Tobias, S. (1994). Interest, Prior Knowledge, and Learning. *Review of Educational Research*, *64*(1), 37–54. https://doi.org/10.3102/00346543064001037

Traub, R. E. (1993). On the Equivalence of the Traits Assessed by Multiple-Choice and Constructed-Response Tests. In W. C. Ward & R. E. Bennett (Eds.), *Construction Versus Choice In Cognitive Measurement. Issues in Constructed Response, Performance Testing, and Portfolio Assessment.* (pp. 29–44). Routledge.

Tucker-Drob, E. M., & Briley, D. A. (2014). Continuity of genetic and environmental influences on cognition across the life span: A meta-analysis of longitudinal twin and

adoption studies. *Psychological Bulletin*, *140*(4), 949–979. https://doi.org/10.1037/a0035893

Tucker-Drob, E. M., & Salthouse, T. A. (2008). Adult age trends in the relations among cognitive abilities. *Psychology and Aging*, *23*(2), 453–460. https://doi.org/10.1037/0882-7974.23.2.453

Tulving, E. (1972). Episodic and Semantic Memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory*. Academic Press.

Tulving, E., & Watkins, M. J. (1973). Continuity between Recall and Recognition. *The American Journal of Psychology*, *86*(4), 739–748. https://doi.org/10.2307/1422081

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin*, *145*(1), 79–139. https://doi.org/10.1037/bul0000176

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2011). Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *The Quarterly Journal of Experimental Psychology*, *64*(3), 447–466. https://doi.org/10.1080/17470218.2010.505292

von Stumm, S. (2013). Investment Traits and Intelligence in Adulthood: Assessment and Associations. *Journal of Individual Differences*, *34*(2), 82–89. https://doi.org/10.1027/1614-0001/a000101

Watrin, L., Geiger, M., Levacher, J., Spinath, B., & Wilhelm, O. (2022). Development and Initial Validation of an Admission Test for Bachelor Psychology Studies. *Frontiers in Education*, *7*, 909818. https://doi.org/10.3389/feduc.2022.909818

Watrin, L., Schroeders, U., & Wilhelm, O. (2022). Structural invariance of declarative knowledge across the adult lifespan. *Psychology and Aging*, *37*(3), 283–297. https://doi.org/10.1037/pag0000660

Watrin, L., Schroeders, U., & Wilhelm, O. (2023). Gc at its boundaries: A cross-national investigation of declarative knowledge. *Learning and Individual Differences*, *102*, 102267. https://doi.org/10.1016/j.lindif.2023.102267

Wilhelm, O. (2005). Measuring Reasoning Ability. In *Handbook of Understanding and Measuring Intelligence* (pp. 373–392). SAGE Publications, Inc. https://doi.org/10.4135/9781452233529.n21

Wilhelm, O., & Kyllonen, P. (2021). To predict the future, consider the past: Revisiting Carroll (1993) as a guide to the future of intelligence research. *Intelligence*, *89*, 101585. https://doi.org/10.1016/j.intell.2021.101585

Wilhelm, O., & Schroeders, U. (2019). Intelligence. In R. Sternberg & J. Funke (Eds.), *The Psychology of Human Thought: An Introduction* (pp. 257–277). Heidelberg University Publishing.

Wilhelm, O., Schulze, R., Schmiedek, F., & Süß, H.-M. (2003). Interindividuelle Unterschiede im typischen intellektuellen Engagement. *Diagnostica*, *49*(2), 49–60. https://doi.org/10.1026//0012-1924.49.2.49

Zell, E., & Krizan, Z. (2014). Do People Have Insight Into Their Abilities? A Metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111–125. https://doi.org/10.1177/1745691613518075

Ziegler, M., Danay, E., Heene, M., Asendorpf, J., & Bühner, M. (2012). Openness, fluid intelligence, and crystallized intelligence: Toward an integrative model. *Journal of Research in Personality*, *46*(2), 173–183. https://doi.org/10.1016/j.jrp.2012.01.002

II Manuscript I

Testing Competing Claims About Overclaiming

The study presented in this manuscript was conceptualized by all co-authors. The applied overclaiming questionnaire was developed by Dr. Selina Weiss, Dr. Diana Steger, and myself. All co-authors contributed to programming of the tasks of this multivariate test battery. The organization and the data collection effort were mainly led by Dr. Selina Weiss and Dr. Diana Steger, and I assisted them in this endeavor. Data analysis regarding the manuscript at hand was mainly conducted by myself under the supervision of all co-authors. I have completed a literature search and drafted the original manuscript. All co-authors reviewed and revised the manuscript under my supervision as first author. The manuscript was accepted for publication in *Intelligence* on 30th June 2020. The manuscript was published online on 11th July 2020.

Republished based on copyright policies from Elsevier; permission conveyed through Copyright Clearance Center, Inc.

Goecke, B., Weiss, S., Steger, D., Schroeders, U., & Wilhelm, O. (2020). Testing competing claims about overclaiming. *Intelligence*, *81*, 1-11. https://doi.org/10.1016/j.intell.2020.101470

Intelligence 81 (2020) 101470

Contents lists available at ScienceDirect

Intelligence

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

Testing competing claims about overclaiming☆

B. Goecke[a,*], S. Weiss[a], D. Steger[a], U. Schroeders[b], O. Wilhelm[a]

[a] Ulm University, Ulm, Germany
[b] University of Kassel, Kassel, Germany

ARTICLE INFO

Keywords:
Overclaiming
Declarative knowledge
Self-reported knowledge
Creativity
Intelligence
Faking

ABSTRACT

Overclaiming has been described as people's tendency to overestimate their cognitive abilities in general and their knowledge in particular. We discuss four different perspectives on the phenomenon of overclaiming that have been proposed in the research literature: Overclaiming as a result of a) self-enhancement tendencies, b) as a cognitive bias (e.g., hindsight bias, memory bias), c) as proxy for cognitive abilities, and d) as sign of creative engagement. Moreover, we discuss two different scoring methods for an OCQ (signal detection theory vs. familiarity ratings). To distinguish between the different viewpoints of what overclaiming is, we juxtaposed overclaiming, as indicated by claiming familiarity with non-existent terms, with fluid and crystallized intelligence, self-reported knowledge, creativity, faking ability, and personality. Overclaiming was measured with a newly comprised overclaiming questionnaire. Results of several latent variable analyses based upon a multivariate study with 298 participants were: First, overclaiming is neither predicted by honesty-humility nor faking ability and therefore reflects something different than mere self-enhancement tendencies. Second, overclaiming is not predicted by crystallized intelligence, but is highly predictive of self-reported knowledge and, thus, not suitable as an index or a proxy for cognitive abilities. Finally, overclaiming is neither related to divergent thinking and originality, and only moderately predicted by self-reported openness creativity from the HEXACO which means that overclaiming does not reflect creative ability. In sum, our results favor an interpretation of overclaiming as a phenomenon that requires more than self-enhancement motivation, in contrast to the claim that was initially proposed in the literature.

1. Introduction

What do you know about the Higgs-Boson? How much do you know about plate tectonics? And what about cholarine? Depending on your knowledge on physical sciences, you may or may not be familiar with the first two topics—the Higgs-Boson and plate tectonics. However, it is impossible that you know anything about cholarine, simply because it does not exist. It is one of the fictitious words (or foils) used in an *Overclaiming Questionnaire* (OCQ; Paulhus, Harms, Bruce, & Lysy, 2003). The phenomenon that people—intentionally or not—overestimate their knowledge of terms, events, concepts, people, etc. that cannot be known because they do not exist is called overclaiming (e.g., Paulhus et al., 2003; Phillips & Clancy, 1972).

Four competing theoretical perspectives on overclaiming can be distinguished. First, Paulhus et al. (2003, p. 899) understood overclaiming as a "trait-like tendency to self-enhance". Second, overclaiming can be seen as a cognitive processing bias (Müller & Moshagen, 2018; Paulhus, 2011). Third, accurately distinguishing between existent and non-existent words on OCQs might be taken as a proxy for declarative knowledge since persons with more knowledge might overclaim less (Paulhus & Harms, 2004). Finally, overclaiming might occur whenever people are willing to encounter, engage with, or elaborate on new concepts or ideas (Dunlop et al., 2017), possibly making overclaiming akin to creativity. Regardless of which perspective one adopts, many studies placed overclaiming into a broader nomological network including faking and cheating (Bing, Kluemper, Kristl Davison, Taylor, & Novicevic, 2011; Feeney & Goffin, 2015; Kam, Risavy, & Perunovic, 2015; Steger et al., 2019a), fluid and crystallized

☆ This research was generously supported by the Bamberg Graduate School of Social Sciences, which is funded by the German Research Foundation (DFG) under the German Excellence Initiative (GSC1024). We thank the numerous research assistants who helped during the data acquisition and human coding.

For the study, we have reported conditions and data exclusion. Decisions surrounding final sample sizes are presented in the sample section. The data is available along with syntax for the statistical analysis at:
https://osf.io/4bnv5/

* Corresponding author at: Institute for Psychology and Pedagogy, Ulm University, Albert-Einstein-Allee 47, Ulm 89081, Germany.
E-mail address: benjamin.goecke@uni-ulm.de (B. Goecke).

https://doi.org/10.1016/j.intell.2020.101470
Received 19 December 2019; Received in revised form 29 June 2020; Accepted 30 June 2020
0160-2896/ © 2020 Elsevier Inc. All rights reserved.

intelligence (Hülür, Wilhelm, & Schipolowski, 2011; Paulhus & Harms, 2004), and the narcissism facet of dark personality (Grosz, Lösch, & Back, 2017; Paulhus et al., 2003; Williams, Paulhus, & Nathanson, 2002). Although overclaiming questionnaires are a frequently used instrument in several disciplines of psychology, our understanding of their underlying nature is still limited.

In the current study, we examine overclaiming as indicated by claiming familiarity with non-existent terms (Phillips & Clancy, 1972) in light of the above-mentioned perspectives by deriving predictions from all viewpoints and testing them within a latent-variable approach. We want to shed light on the relation of overclaiming with declarative knowledge, self-reported knowledge, creative ability, honesty-humility and openness creativity, faking ability, and fluid intelligence.

1.1. Scoring methods of overclaiming: Raw responses vs. signal detection indices

A number of researchers worked on different approaches to assess (e.g., Ackerman & Ellingsen, 2014; Ziegler, Kemper, & Rammstedt, 2013) and score overclaiming (Hülür et al., 2011; Paulhus & Petrusic, 2007). The basic idea of using foils for studying familiarity claims can be traced back to Raubenheimer (1925), who laid the foundation of the *Overclaiming Technique* (OCT). Afterwards, this method had long been forgotten until it was revived in consumer behavior research (Phillips & Clancy, 1972). The OCT finally made its way to individual differences research in an attempt to use the technique for studying self-enhancement tendencies independent of ability (Paulhus et al., 2003). In the original OCQ (OCQ-150; Paulhus et al., 2003), participants were asked to rate their familiarity with 150 items by rating them on a scale ranging from "*never heard of it*" to "*know it very well*". Since then, the OCT and a variety of OCQs were developed and used in psychological research (e.g., Atir, Rosenzweig, & Dunning, 2015; Dunlop et al., 2017, 2019; Mesmer-Magnus, Viswesvaran, Deshpande, & Joseph, 2006; Pennycook & Rand, 2019; Ziegler et al., 2013). Although alternative assessment methods of overclaiming are interesting, we will not further discuss them here and will instead focus on competing scoring methods of conventional OCQs.

With respect to the methodology tied to the OCT, it is necessary to distinguish several terms and statistical indices. An OCQ consists of two different types of items, namely reals and foils. Whereas reals refer to items that are existent terms, concepts, or words, foils are non-existent as terms, concepts, or words. Thus, reals and foils belong to the item-side of an OCQ. Based on participants' answers to these items, several person indices can be computed. On the one hand, the familiarity rating of reals and foils may be read precisely as such: as ordinary raw responses on a rating scale for familiarity. Asking participants to rate their familiarity with existent terms, concepts, or words, probes their self-reported knowledge (e.g., Bing et al., 2011; Fell, König, Jung, Sorg, & Ziegler, 2019; Hülür et al., 2011), whereas familiarity ratings with items that cannot be familiar indicate overclaiming (Bing et al., 2011; Phillips & Clancy, 1972; Randall & Fernandes, 1991). This scoring of an OCQ is straightforward and economical, as it only requires the independent computation of mean familiarity ratings for reals and foils to express a person's standing on self-reported knowledge and overclaiming, respectively.

On the other hand, a signal detection approach is commonly used and often recommended to analyze OCQ data (e.g., Paulhus, 2011; Paulhus et al., 2003). The two most prominent indices that are computed within the framework of signal detection are the accuracy index *d'* (d prime, also called *OC accuracy*) and the index for response bias *c* (criterion location, also called *OC bias*). *d'* is thought to be indicative of a person's sensitivity to accurately categorize reals as existing and foils as non-existing; *c* can be understood as an indicator of a person's "yes-tendency", indicating a general tendency to claim familiarity with the items of an OCQ (MacMillan & Creelman, 2005; Paulhus et al., 2003).

A number of approaches have been developed to obtain both scores (e.g., Stanislaw & Todorov, 1999; Wixted, 2020). All approaches have in common that *d'* and *c* rely on the same information for computation, namely the hit rate and the false alarm rate from a binary classification table. The hit rate is usually computed as the proportion of correct familiarity ratings for reals and the false alarm rate is computed as the proportion of false familiarity ratings for foils. Importantly, *d'* and *c* rely on ratings for both reals and foils and are logically and computationally dependent. This is problematic in a statistical sense, because psychometrically the accuracy index should be unaffected by the response bias to be interpretable. Yet, *d'* is only unaffected by *c* if the distributions of familiarity rating of signal trials (which translates to the hit rate) and familiarity ratings of noise trials (which corresponds to the false alarm rate) are normally distributed, have equal variances (Kellen & Klauer, 2018; Stanislaw & Todorov, 1999), and are uncorrelated. However, these assumptions are usually not met (e.g., DeCarlo, 2010; MacMillan & Creelman, 2005; Mickes, Wixted, & Wais, 2007; Selker, van den Bergh, Criss, & Wagenmakers, 2019; Stanislaw & Todorov, 1999; Starns & Ratcliff, 2014; Swets, 1986).

Besides these statistical considerations, both indices are difficult to grasp in terms of the underlying cognitive processes and the conceptual link between manifest indicators and latent constructs. While the interpretation of responses to reals and foils as familiarity ratings in a latent variable approach is straightforward, this interpretation is not clear in signal detection theory. Two processes—the correct identification of reals and the correct rejection of foils—are intertwined, so that a lower score in OC bias can be achieved by a better hit rate, a lower false alarm rate, or both.

Taken together, indices of signal detection theory are more problematic and less straightforward indicators to exhaust responses from an overclaiming questionnaire than commonly assumed. The *d'* and *c* indices are confounded by virtue of being computed based on the same underlying information (hit rate and false alarm rate) and the information captured in ratings does not meet statistical assumptions for computing adequate person estimates for bias and accuracy. Due to these reasons, we do not pursue the frequently used signal detection scoring approach and rather rely on the information provided by mean familiarity ratings of reals and foils, respectively. We use these ratings as manifest variables in a structural equation model to examine individual differences in self-reported knowledge and overclaiming. However, we acknowledge the widespread use of signal detection analysis for OCQs and therefore provide a set of results and a description of the corresponding formulas (see supplementary material: SM Fig. 1, SM Table 2, SM Table 3).

1.2. What drives overclaiming behavior

In the following, we discuss four theoretical perspectives on the phenomenon of overclaiming as measured by an OCQ. That is, (a) overclaiming reflects self-enhancement tendencies, (b) overclaiming is a product of a cognitive processing bias, (c) rating familiarity for reals and foils on an OCQ is an index of cognitive abilities, and (d) overclaiming is a type of creative engagement. An overview of the theoretically-derived predictions made by the four theoretical accounts is given in Table 1 for the above-discussed mean familiarity ratings for reals (SRK) and foils (OC) on the one hand, and for the conventional signal detection indices (OC bias and OC accuracy) on the other hand.

From the first perspective, overclaiming behavior is understood as reflecting a overarching self-enhancement tendency, which describes the tendency to present oneself in a desirable way (e.g., when completing self-reports). In line with this notion, Paulhus et al. (2003) found convergent validity of the OC bias index with established measures of self-enhancement such as the *Narcissistic Personality Inventory* ($r = 0.35$; Raskin & Hall, 1979) and the *Self-Deceptive Enhancement Scale* ($r = 0.30$; Paulhus & Reid, 1991). Paulhus et al. (2003) concluded that the OCT is useful to assess the proneness to self-enhance, which would be a useful instrument in various contexts that rely on self-report

Table 1
Synopsis of Theoretical Perspectives on Overclaiming and Corresponding Predictions.

Index	Overclaiming (OC)	Self-Reported Knowledge (SRK)	OC bias (*c*)	OC accuracy (*d'*)
If responses on OCQ are influenced by…				
Self-enhancement tendencies	+	+	+	−
Cognitive Biases (e.g., hindsight bias, memory bias)	+	+	+	−
Cognitive Abilities (i.e., knowledge)	−	+	−	+
Creative Engagement	+	+	+	−

Note. The table describes which index of an OCQ should show a positive (+) or negative (−) relation to constructs that are reflective for the theoretical assumption in question (e.g., a *gc* measure for the perspective that an OCQ is reflective for cognitive abilities). OC and SRK are based on familiarity ratings of foils and reals, respectively. OC bias and OC accuracy are both calculated based on the hit rate and the false alarm rate.

measures (e.g., personnel selection), as these frequently suffer from various biases (Van Vaerenbergh & Thomas, 2013; Ziegler, 2015). Generally, engaging in overclaiming has been described as a potentially economical marker for positivity biases or self-enhancement tendencies (e.g, Bensch, Paulhus, Stankov, & Ziegler, 2017; Dunlop et al., 2019; Paulhus et al., 2003). It has been considered as a type of remedy and an efficient mean of assessing individual differences in the tendency to systematically misrepresent oneself in self-reports (Ludeke & Makransky, 2016; Paulhus et al., 2003; Paulhus & Harms, 2004). In this sense, overclaiming might reflect people's tendency to present themselves in a socially desirable way, that is, to intentionally exaggerate self-reports on desirable attributes, which is also termed *positive impression management* (Paulhus, 1984; Paulhus & Reid, 1991). In the same vein, subjects may claim to have more knowledge than they actually possess because they unintentionally overstate or overrate their abilities. This form of overclaiming might be labeled *self-deceptive enhancement* (Paulhus & Reid, 1991). The distinction between intentionally and unintentionally overclaiming is interesting, as scales measuring impression management and self-deceptive enhancement have been found to be almost unrelated (Paulhus, 2002), indicating a substantial difference between consciously and subconsciously misrepresenting oneself. Accordingly, overclaiming could be seen as a way to positively misrepresent oneself, which puts the concept in relation to the personality factor of honesty-humility (Ashton & Lee, 2001), as well as faking (i.e., faking good, Ziegler, 2011). Although both honesty-humility and faking are not direct measures of self-enhancement, overclaiming, honesty-humility, and faking share the idea of a persons' misrepresentation. However, faking and dishonesty is more of a motivated deception, while self-enhancement overlaps with positive self-beliefs such as self-esteem and confidence.

Regarding the empirical evidence, impression management could be predicted by honesty-humility (de Vries, Zettler, & Hilbig, 2014). This relationship makes sense, because describing oneself (deliberately) in a more positive manner should correspond to being less honest and less humble. The same logic can be applied to the association of overclaiming with honesty-humility, although recent evidence shows no relation between different indices of overclaiming and honesty-humility (OC bias: Dunlop et al., 2019; Müller & Moshagen, 2019; mean familiarity ratings of foils: Steger, Schroeders, & Wilhelm, 2020), thus suggesting overclaiming corresponds to unintentionally overrating one's abilities.

Apart from honesty-humility, faking was repeatedly described as a deliberate positive self-presentation or an intentional bias (e.g., Dunlop et al., 2019; Geiger, Olderbak, Sauter, & Wilhelm, 2018; Ziegler, 2011), and the degree to which it met requirements should be considered an ability (Geiger et al., 2018). In line with this idea and given theoretical accounts of faking ability (e.g., Ziegler, 2011), it should be considered that the setting in which an OCQ is administered (low-stakes vs. high-stakes) might play an important role for the validity of overclaiming as a potential marker of deliberate positive self-representations such as faking (Dunlop et al., 2019). More precisely, the evidence that the OC bias index is predictive of faking in low-stakes settings is rather weak (Dunlop et al., 2019; Feeney & Goffin, 2015; Ludeke & Makransky, 2016; Müller & Moshagen, 2019), higher in high-stakes setting (e.g., Bing et al., 2011; Dunlop et al., 2019), but still not ubiquitous (Bensch, Maaß, Greiff, Horstmann, & Ziegler, 2019).

From the second perspective, engaging overclaiming can also be seen as a sort of cognitive processing bias. It was argued that due to the design of the OCQ—around 20% foils embedded in about 80% reals—the familiarity of an item might be overrated, resulting in a memory bias (Müller & Moshagen, 2018; Paulhus, 2011). That is, foils are typically produced and presented in a manner that makes them plausible and sometimes even orthographically or linguistically similar to reals (Ackerman & Ellingsen, 2014). To this end, the item construction rationale of an OCQ might lead to increased processing fluency within persons while evaluating foils. Due to their resemblance with real words, a false feeling of familiarity can lead to a false feeling of knowing (Müller & Moshagen, 2018). In turn, this accessibility would result in a high familiarity rating of non-existing items (Dunlop et al., 2017). This kind of spurious sense-making and the apparent misattribution of fluency for real familiarity are processes that are also linked to reconstruction biases of the well-known hindsight bias, which has been shown to explain some of the variance in a variant of the OC bias parameter (Müller & Moshagen, 2018). This perspective contributes to the understanding of overclaiming by emphasizing the subconscious rather than the conscious account of it. However, contrary to seeing overclaiming as subconscious and self-deceptive behavior, the explanation for overclaiming from this point of view focuses on issues emerging from the typical method of measurement, namely questionnaires, and its elements, namely single items. That is, overclaiming occurs inadvertently, due to shallow characteristics of foils, and not due to any self-enhancement motivations.

Third, overclaiming has also been juxtaposed with several cognitive abilities such as fluid intelligence (*gf*) and crystallized intelligence (*gc*)—both self-reported and objectively measured. People generally tend to overestimate their own abilities in self-reports, and their own knowledge in particular (Kruger & Dunning, 1999). Correlations between self-reported abilities and objective tests of cognitive abilities are only moderate, even when persons rating themselves received training in self-estimation (Jacobs & Roodenburg, 2014). This finding is further

supported by meta-analytically estimated correlations between self-reports and objective tests of cognitive abilities, which are also only moderate (r = 0.33, Freund & Kasten, 2012; ρ = 0.29, Mabe & West, 1982), demonstrating a mismatch between persons' self-perceived and actual abilities. Nonetheless, it has been argued that the OC accuracy index might be used as a proxy for *gc* (Paulhus & Harms, 2004), despite the fact that the moderate correlations between the OC accuracy index and *gc*—varying between r = 0.24 and 0.50 (Bensch et al., 2017; Hülür et al., 2011; Hülür, Wilhelm, & Schipolowski, 2014; Paulhus & Harms, 2004; Ziegler et al., 2013)—have raised doubts about this claim. In sum, the signal detection scored familiarity ratings (i.e., OC accuracy) on OCQs were described as a "practical measure of global cognitive ability" (β = 0.51; Paulhus & Harms, 2004, p. 311), but the empirical evidence paints a more nuanced picture.

Finally, Hülür et al. (2011, 2014) linked overclaiming with intellectual investment traits such as typical intellectual engagement, self-motivated cognition, and the personality facet of openness to new ideas. Given the strong conceptual and empirical overlap between these concepts (Mussel, 2013), one could argue that overclaiming behavior is a byproduct of open-minded people who value new ideas, concepts, and terms; these people tend to overrate their familiarity with novel items. This perspective is supported by results showing that typical intellectual engagement and openness—as measured by the five-factor model (McCrae & Costa, 1989)—are positively related to overclaiming (Ackerman & Ellingsen, 2014). Such results might indicate that subjects scoring higher in openness encounter unknown terms in a less skeptical or rational manner but are willing to develop ad hoc ideas of what an unfamiliar term might mean. In the alternative six-factor model of personality (HEXACO; Ashton & Lee, 2001), the openness factor (originally termed intellect/imagination) includes the facets of aesthetic appreciation, inquisitiveness, creativity, and unconventionality. Both openness factors (from the HEXACO and the five-factor model) are highly correlated and reported as mostly congruent (r = 0.71; Moshagen, Hilbig, & Zettler, 2014). Especially, the openness factor based on the HEXACO includes items that might trigger endorsement from participants who feel an attraction towards foils, because these items emphasize the experience of the unknown. This idea is supported by medium correlations between OC bias and openness in the HEXACO taxonomy (r = 0.34, Dunlop et al., 2017). Findings relating short scales of the openness factor based on the five-factor model with OC bias were inconclusive (r = 0.02, Bensch et al., 2017; r = 0.25, Williams et al., 2002). The HEXACO factor of openness has also been predictive of creativity (β = 0.55; Silvia, Kaufman, Reiter-Palmon, & Wigert, 2011), for the reason that more creative persons may be prone to overclaim as they are more willing to encounter, engage in, and elaborate on unknown and unfamiliar ideas, concepts, or terms. Bearing these considerations in mind, examining the relationship among the creativity facet of openness (as proposed by the HEXACO), creativity as an ability, and overclaiming might further our understanding on what possibly drives overclaiming behavior.

1.3. Research aims

Given the competing perspectives on overclaiming and its potential utility for practitioners, it seems pivotal to examine overclaiming more broadly within a comprehensive nomological network in order to further elaborate on its very nature. The essential goal of our study was to provide a test of the accounts outlined above that understand overclaiming as an interesting phenomenon that is—from our point of view—driven by more than a cognitive processing bias and should therefore be considered from a variety of viewpoints.

a. First, we present a measurement model of overclaiming and self-reported knowledge based on mean familiarity ratings.
b. Then, according to the first perspective that overclaiming behavior is part of the nomological net of a persons' tendency to misrepresent, we test the prediction that overclaiming and self-reported knowledge are partially accounted for by honesty-humility and faking ability, because the conscious tendency to positively misrepresent oneself (as presumably measured with the OCQ) should correspond to individual differences in honesty-humility and faking ability—-where less honest/humble individuals engage more in overclaiming, and individuals better in faking also tend to overclaim more.
c. In addition, concerning the idea that overclaiming behavior might reflect cognitive abilities, we examine the amount of variance in overclaiming and self-reported knowledge that could be explained by *gc* and *gf* (see also Hülür et al., 2011).
d. Finally, we try to put the fourth perspective to a critical test. We do so by regressing individual differences in overclaiming and self-reported knowledge on individual differences in creative abilities and the creativity items of the openness scale. According to the fourth perspective, a substantial amount of variance in overclaiming and self-reported knowledge should be explainable by creativity (i.e., divergent thinking and originality) and openness (depicted by creativity indicators).

2. Method

2.1. Participants and procedure

The current examination was part of a larger, multi-center study assessing creativity (Weiss et al., 2020). The analyses rely on a subsample of 298 participants after data cleaning (for more information see Weiss et al. (2020)). The participants completed seven hours of (mostly) computerized testing and were financially compensated. Five hours of testing were done in a laboratory setting, while nearly two hours of testing could be done from home. All participants provided informed consent for the study and a local ethics committee approved the data collection. The majority of the participants reported German as their first language (N = 280) and had a high school diploma (N = 278). Non-native speakers were sufficiently fluent in German to complete all measures. Seventy-two percent of the sample were female; mean age was M = 24.5 years (SD = 5.1). We restrict our description of indicators to the measures pertinent to the present research questions.

2.2. Measures

2.2.1. Overclaiming questionnaire

We used a newly-compiled set of 149 items to measure overclaiming with 120 reals and 29 foils in five broad domains (Natural Sciences, Life Sciences, Social Sciences, Humanities, and Pop Culture), in line with findings on the dimensionality of knowledge (Steger, Schroeders, & Wilhelm, 2019). All items were selected based on an overall collection of n = 1033 items (partly developed by the authors, partly based on Hülür et al., 2011) that covered a broad range of content, for example, for natural sciences items about physics, chemistry, computer science. Prior to item selection, 12 raters evaluated the quality and the dissimilarity of the foils to existing words as well as the difficulty of the reals for half of the items each. The quality of foils was rated on a scale from one (low quality) to five (high quality) based on their surface similarity towards reals (e.g., low similarity in spelling indicates a high quality) and their believeability (e.g., poor plausibility indicated low quality). The difficulty of reals was rated in years of education (on a scale from one to 20, including tertiary education) necessary to solve

the item with a probability of 0.80. For the final item set, we selected reals and foils with high difficulties and of sufficient quality. The aggregated data of these ratings for the selected item set, as well as the original mean familiarity ratings of the items that were selected from the OCQ of Hülür et al. (2011) are presented in the supplementary material (SM Table 1). Regarding the response scale of the OCQ, we chose a format with fully-labeled response options (e.g., Weng, 2004) that has been successfully used in the past (Hülür et al., 2011). Participants had to indicate their familiarity to words on a 5-point scale (1 = *Never heard of it before*, 2 = *Heard of it before, but cannot describe it exactly*, 3 = *I could describe it roughly*, 4 = *I could describe it relatively precise*, and 5 = *I could give an exact description*). For the complete questionnaire, see the online repository.

2.2.2. Fluid and crystallized intelligence

Gf was measured with the verbal and the figural subtests of the *Berlin Test of Fluid and Crystallized Intelligence* (BEFKI; Wilhelm, Schroeders, & Schipolowski, 2014). The verbal aspect of *gf* was assessed via relational reasoning, that is, participants had to derive logically-valid conclusions based on a set of given premises. The figural aspect of *gf* was composed of a sequence of geometric drawings that changed their shading and form according to certain rules. Participants had to infer these rules and to select missing drawings out of several response alternatives. Two items of the verbal scale had to be removed from further analysis due to extreme difficulties. *Gc* was measured with two parallel test forms compiled of 102 randomly-assigned items each from a larger pool of declarative knowledge items (Steger, Schroeders, & Wilhelm, 2019, 2020). Also, the broad content domains of *gc* matched the domains of the overclaiming questionnaire.

2.2.3. Creativity

The test battery included a variety of verbal and figural divergent thinking tests (Carroll, 1993), as they are used for the measurement of creative ability along two domains: fluency (quantity of results in a given time) and originality (quality of results). The battery included four retrieval fluency tests: the similar attributes (SA including six items) and the inventing names (IN including 18 items) tests were both adapted from the "Verbaler Kreativitätstest" [verbal creativity test] (VKT; Schoppe, 1975), while the retrieval fluency (RF including six items) and the figural fluency (FF including four items) test were used from the Berlin Structure-of-Intelligence test (Jäger, Süß, & Beauducel, 1997). Originality was assessed with two tests: combining objects (CO, including 12 items) (French, Ekstrom, & Price, 1963) and inventing nicknames (NI, including nine items) (VKT; Schoppe, 1975). Divergent thinking items are usually open-ended and require human rating. All answers were counted (fluency) or scored (originality) by three independent human raters. The originality tests were coded regarding the creativity of an answer on a five-point Likert scale (Amabile, 1982; Silvia et al., 2008), resulting in good inter-rater reliability (Weiss et al., 2020). The measurement model was specified with test-scores (i.e., mean across all rated items) for each of the mentioned tests as indicators. Two creativity tests measuring originality load on a nested factor, in addition to an overarching divergent thinking factor. For a more detailed description of the tests, their psychometric properties, and the measurement model of divergent thinking, please see Weiss et al. (2020).

2.2.4. Personality

Personality was measured with the German 60-item version of the HEXACO (Moshagen et al., 2014), covering the personality factors of *Honesty-Humility*, *Emotional Stability*, *Extraversion*, *Agreeableness*, *Conscientiousness,* and *Openness,* with 10 items each. Participants had to indicate their agreement with statements on a five-point scale ranging from *"strongly disagree" to "strongly agree"* (e.g., *"I'm interested in learning about the history and politics of other countries"* for openness). For modeling honesty-humility (see Fig. 2), we used three parcels with similar mean values. Based on the theoretical considerations in the introduction, the openness factor in Fig. 4 only includes the three items of the openness scale that tap creativity (Ashton & Lee, 2009).

2.2.5. Faking

Faking ability was measured using the paradigm presented in Geiger et al. (2018) with items from the Work Style Questionnaire (WSQ), a self-report measure originally developed for O*NET (Borman, Kubisiak, & Schneider, 1999). The questionnaire consists of 16 items defining various aspects of a typical work style (e.g. Achievement/Effort: "Establishing and maintaining personally challenging achievement goals and exerting effort towards mastering tasks"). Participants were asked to select a response option from a five-point Likert scale ranging from "Not at all like me" to "Extremely like me" for all 16 items for five different jobs with distinct work style patterns (pilot, talk show host, software engineer, tour guide, insurance seller). They were instructed to select response options in a way to maximize their chances of being hired for the five jobs in a personality task-style fashion ("Please rate how much this characteristic describes you"). For scoring, we applied the profile similarity metric shape as reported in Geiger et al. (2018), i.e. the participants' responses were correlated to expert ratings of the jobs work styles from O*NET. Slightly alienating from Geiger et al. (2018), we rounded the optimal profile scores (rational numbers) in order to compute Spearman rank-order correlations per participant, due to administering a discrete five-point Likert scale (natural numbers) in our study. These correlations were used as indicators for a general latent model of faking (Geiger et al., 2018). For validation purposes of the faking factor, we report a two-factor CFA with *gc* in the supplementary material (SM Fig. 2), which shows a correlation between the factor of faking and *gc* of $r = 0.41$ (similar to results found by Geiger et al., 2018).

2.3. Statistical analysis

The main analyses were conducted within the framework of *structural equation modeling* (SEM) with R (R Core Team, 2018) and the R package *lavaan* (Rosseel, 2012). The models presented in this paper are based on indicators of mean familiarity ratings of reals and foils. In order to circumvent possible problematic distributions of the variables (i.e., non-normally distributed indicators), the reported models relied on a robust maximum likelihood estimator where necessary. To reduce model complexity, we used item parcels as indicators after establishing

Table 2.
Descriptive statistics for familiarity ratings, including the means, standard deviations, skews, and Kurtosis for each Pparcel.

			Reals			Foils		
Parcel	no. reals	no. foils	*M (SD)*	Skew	Kurtosis	*M (SD)*	Skew	Kurtosis
1	25	8	2.74 (0.56)	0.20	−0.12	1.55 (0.43)	1.33	2.38
2	29	6	2.68 (0.50)	0.29	0.19	1.70 (0.57)	1.02	1.26
3	24	6	2.67 (0.51)	0.23	−0.25	1.60 (0.47)	1.50	2.64
4	22	5	2.62 (0.49)	0.25	−0.17	1.40 (0.44)	1.71	3.75
5	21	4	2.88 (0.56)	0.23	−0.05	1.47 (0.51)	1.63	3.26

Table 3
Measurement models for overclaiming, self-reported knowledge, crystallized and fluid intelligence, and creativity.

Measurement Model		*n* (Parcels)	*n*(Persons)	χ^2	*df*	CFI	RMSEA [90% CI]		SRMR	ω
1	OC	5	297	4.70	5	1.00	0.000	[0.000; 0.067]	0.012	0.90
2	SRK	5	297	14.19	5	0.994	0.079	[0.034; 0.127]	0.007	0.97
3	Gc	5	293	6.34	5	0.996	0.030	[0.000; 0.091]	0.021	0.78
4	Gff / Gfv	8	297	18.48	19	1.00	0.000	[0.000; 0.050]	0.028	0.71 / 0.69
5	Faking Ability	5	295	8.37	5	0.936	0.048	[0.000; 0.103]	0.032	0.46
6	DT / Originality*	6	298	5.72	8	1.00	0.000	[0.000; 0.056]	0.017	0.77 / 0.28

Note. OC = Overclaiming; SRK = Self-Reported Knowledge; Gc = Crystallized Intelligence; Gfv = Fluid Intelligence verbal; Gff = Fluid Intelligence figural. DT (divergent thinking) is a general factor, while originality is nested in DT. ω = McDonald's Omega (McDonald, 1999).

the unidimensionality of the scales (Little, Cunningham, Shahar, & Widaman, 2002). In more detail, we parceled the indicators for overclaiming, self-reported knowledge, and knowledge items heterogeneously so that each parcel was stratified across all broad content domains (Cole, Perkins, & Zelkowitz, 2016; Steger et al., 2019). Furthermore, the intelligence scales were parceled with equal average difficulty. We used *full information maximum likelihood* estimation under the assumption of missing at random (i.e., missingness due to program failures: less than 2% of cell entries were missing) to combine missing data and parameter estimation in a single step (Enders, 2010). The following fit indices were considered as indication for good model fit: CFI (*Comparative Fit Index*) ≥ 0.95, RMSEA (*Root Mean Square Error of Approximation*) ≤ 0.06, and the SRMR (*Standardized Root Mean Square Residual*) ≤ 0.08 (Hu & Bentler, 1999). Prior to modeling, all variables were *z*-standardized. To make all analyses transparent and reproducible, we provide all material necessary to replicate the main findings in an online repository: https://osf.io/4bnv5/

3. Results

In Table 2, we report means and standard deviations of the mean familiarity ratings of the parcels that equally include the broad knowledge domains (Steger et al., 2019). The range of the descriptive statistics of the obtained parcels are similar to the descriptive statistics of other OCQs used the literature (e.g., Dunlop et al., 2019; Hülür et al., 2011).

3.1. Measurement models

To test for unidimensionality of the scales, we first report the fit indices for all individual measurement models (see Table 3). For all measurement models, fit indices were above the aforementioned cut-off values, indicating good to excellent model fit. However, the RMSEA of the measurement model for self-reported knowledge (Model 2) was insufficient, while CFI and SRMR both indicated that model fit was good. This deviating result concerning model fit might be due to the fact that the RMSEA tends to be over-sensitive for models with low degrees of freedom (Kenny, Kaniskan, & McCoach, 2015). For faking ability, the CFI was below the threshold, while the RMSEA and the SRMR indicated a good model fit. We do not report fit indices for the measurement models of honesty-humility and openness, as these models only contained three indicators and were therefore exactly identified.

In addition to the unidimensionality of the scales, we report a latent regression model, where overclaiming (i.e., a latent variable capturing the communality of five mean familiarity rating parcels for foils) predicts self-reported knowledge (i.e., a latent variable capturing the communality of five mean familiarity rating parcels for reals), as this model of the OCQ is the basis for all subsequent models (Fig. 1; see SM Fig. 3 for unstandardized parameters). We have selected a model in which overclaiming predicts self-reported knowledge. Overclaiming accounted for 50.5% of the variance in self-reported knowledge in our sample.

3.2. Overclaiming as a self-enhancement tendency

In line with the first perspective, that overclaiming is a manifestation of an overarching tendency to self-enhance, overclaiming and self-reported knowledge should be predicted by self-reported honesty-humility, as well as the ability to fake (Fig. 2, see SM Fig. 4 for unstandardized parameters). However, honesty-humility did neither significantly predict self-reported knowledge nor overclaiming, whereas faking predicted self-reported knowledge, but not overclaiming. Thus, individual differences in honesty-humility and faking did not explain interindividual differences in overclaiming (R^2 < 1%). For self-reported knowledge, a total of 58% in the variance was explained by all three predictors, however, it was only partly driven by the individual differences in faking ability. Given the empirical overlap between faking ability and *gc* (e.g., Geiger et al., 2018), the explanatory power of faking ability for incrementally predicting self-reported knowledge is vanishingly low: extending the model with a factor of *gc*, the explanatory power of faking ability for self-reported knowledge disappears (β = 0.06; $\chi^2(220)$ = 332.45, CFI = 0.969, RMSEA = 0.041, SRMR = 0.049).

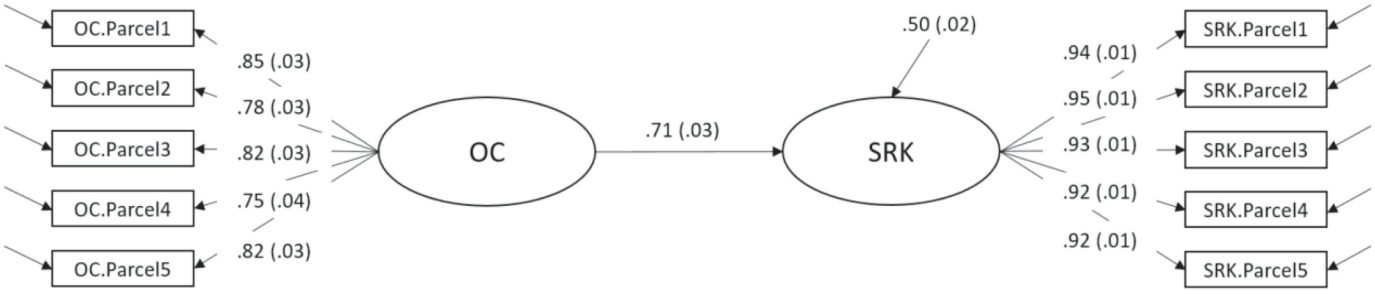


Fig. 1. Structural Model of Overclaiming (OC) and Self-Reported Knowledge (SRK).
Note. Structural model of overclaiming (OC) and self-reported knowledge (SRK). All parameters are standardized. n = 298; $\chi^2(34)$ = 90.87, CFI = 0.978, RMSEA = 0.075, SRMR = 0.038. The model including its unstandardized parameter estimates can be found in the supplementary material (SM Fig. 3).

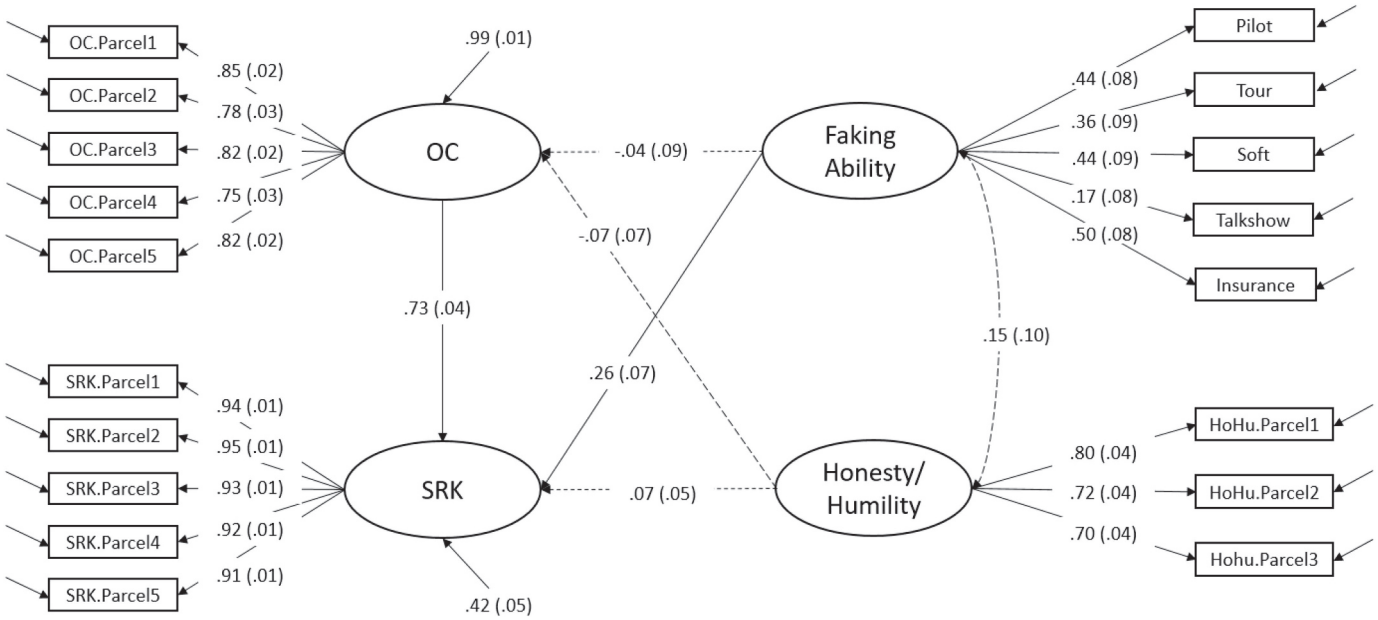


Fig. 2. Structural Model of Overclaiming (OC), Self-Reported Knowledge (SRK), Honesty-Humility, and Faking Ability.
Note. Structural model of overclaiming (OC), self-reported knowledge (SRK), honesty-humility, and faking ability. All parameters are standardized; non-significant coefficients are displayed with dashed lines ($p < .05$). $n = 298$; $\chi^2(129) = 215.64$, CFI = 0.972, RMSEA = 0.047, SRMR = 0.049. The model including its unstandardized parameter estimates can be found in the supplementary material (SM Fig. 4).

3.3. Overclaiming as an index of cognitive ability

Next, we juxtaposed overclaiming and self-reported knowledge with *gc* and *gf* to examine whether cognitive abilities account for the variance in overclaiming and self-reported knowledge (Fig. 3, see SM Fig. 5 for unstandardized parameters). Here, the variance of overclaiming was not explained by *gc* or *gf* at all, whereas 73% of the variance of self-reported knowledge was explained by overclaiming and *gc*. As expected, self-reported knowledge was largely predicted by overclaiming and only moderately predicted by actual *gc*. Although the regression weight between *gc* and self-reported knowledge is in line with similar findings in the literature (e.g. Freund & Kasten, 2012), these results generally and independent of the scoring approach discourage the use of overclaiming scales or self-report questionnaires as proxy measures of broadly defined cognitive abilities such as *gc*. However, because the OCQ can be transformed into simple word/non-word decision task, the mean familiarity ratings with reals correspond to a substantial degree with vocabulary tests (e.g., Ackerman and Ellingsen (2014) found a correlation of $r = 0.71$ between self-claimed vocabulary knowledge and a vocabulary score, and a correlation of $r = 0.43$ between self-rated knowledge and the false-alarm rate). The relations with *gf* are so small that clearly self-reported knowledge cannot possibly serve as a proxy for *gf*.

Additionally, in the structural model modeling both *gc* and *gf*, we tested whether *gc* moderates the link between SRK and OC in a latent

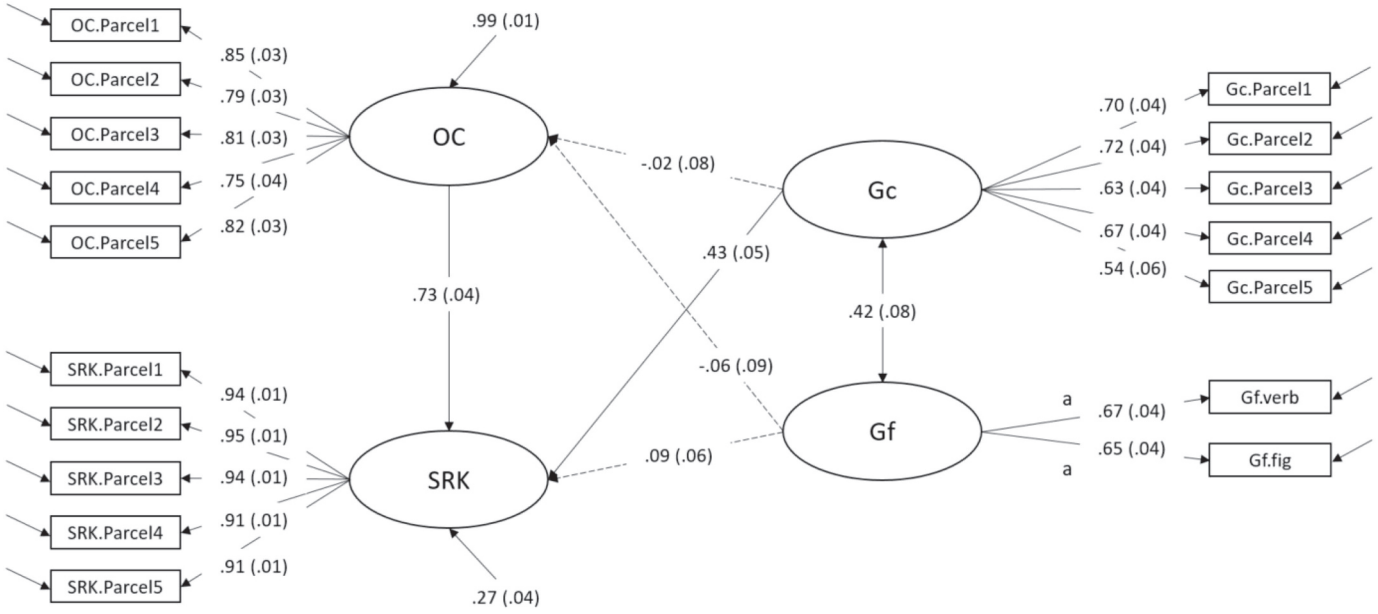


Fig. 3. Structural Model of Overclaiming (OC), Self-Reported Knowledge (SRK), and Knowledge (Gc), and Reasoning Ability (Gf).
Note. All parameters are standardized; non-significant coefficients are displayed with dashed lines ($p < .05$). $n = 298$; $\chi^2(114) = 199.78$, CFI = 0.975, RMSEA = 0.050, SRMR = 0.040. The model including its unstandardized parameter estimates can be found in the supplementary material (SM Fig. 5).

variable model. To do so, we first extracted the factor scores of OC and *gc* in order to compute an interaction variable based on the product of the factor scores of OC and *gc*. This new variable was then entered into a latent regression model, where SRK was predicted by OC, *gc*, and the standardized interaction term. The model fitted the data: $\chi^2(102) = 203.15$, CFI = 0.972, RMSEA = 0.058, SRMR = 0.048. While the regression parameters were similar to the model above for the single predictors ($\beta_{OC} = 0.72$; $\beta_{gc} = 0.46$, $p < .001$), the regression parameter of the interaction term was zero ($\beta_{OC*gc} = 0.00$, $p = .99$). This indicates that the association between SRK and OC is independent from the real knowledge of persons.

3.4. Overclaiming as a form of creative engagement

Finally, we modeled overclaiming and self-reported knowledge together with openness-creativity, divergent thinking, and originality. According to this perspective, overclaiming is a byproduct of creative engagement. In this model, overclaiming and self-reported knowledge were predicted by divergent thinking, originality, and a factor for creativity self-reports from the openness facet of the HEXACO (Fig. 4, see SM Fig. 6 for unstandardized parameters). Overclaiming was significantly predicted by openness-creativity and self-reported knowledge was significantly predicted by divergent thinking. The predictors of creativity and openness were able to explain 10% of the variance of overclaiming and a total of 59% of the variance of self-reported knowledge was explained by all predictors.

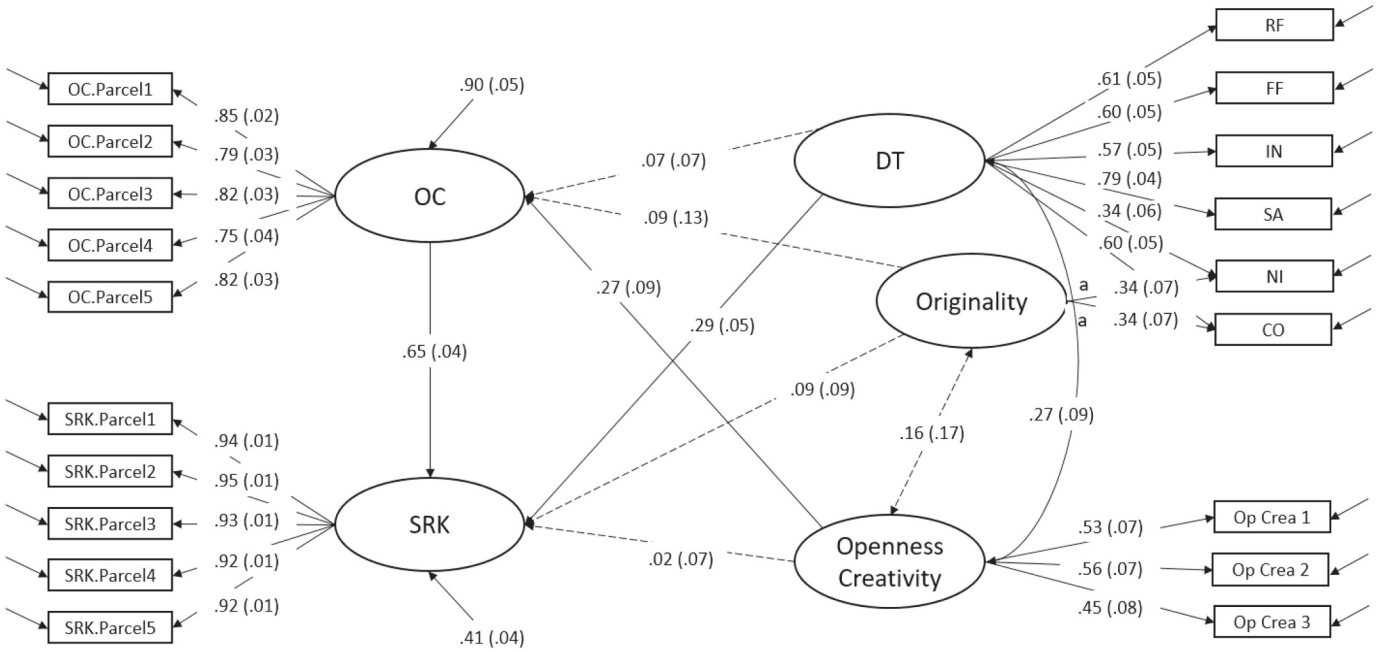


Fig. 4. Structural Model of Overclaiming (OC), Self-Reported Knowledge (SRK), and Divergent Thinking (DT), Originality, and Openness.
Note. RF = Retrieval Fluency, FF = Figural Fluency, IN = Inventing Names, SA = Similar Attributes, NI = Nicknames, CO = Combining Objects. All parameters are standardized; non-significant coefficients are displayed with dashed lines ($p < .05$). $n = 298$; $\chi^2(142) = 225.38$, CFI = 0.975, RMSEA = 0.044, SRMR = 0.044. The model including its unstandardized parameter estimates can be found in the supplementary material (SM Fig. 6).

4. Discussion

The concept of overclaiming as a disposition is still vague and previous interpretations vary broadly, although there is some agreement in the literature as to the potential use of OCQs due to their efficient measurement properties (Dunlop et al., 2019). For our study, we distinguished between two fractions of an OCQ, namely, overclaiming (that is, mean familiarity ratings of foils) and self-reported knowledge (that is, mean familiarity ratings of reals). We linked both constructs with four perspectives on of what responses on OCQs might be reflective: self-enhancement tendencies (Paulhus et al., 2003), a cognitive processing bias due to the plausibility of foils (Müller & Moshagen, 2018; Paulhus, 2011), a "practical measure of global cognitive ability" (Paulhus & Harms, 2004, p. 311), and a form of creative engagement. The purpose of the present study was to critically test these interpretations and to strengthen the evidence with respect to the relation between overclaiming, self-reported knowledge, openness, and creative abilities. To understand individual differences in overclaiming and their antecedents, covariates, and consequences, we compiled an OCQ out of an existing measurement and newly-developed items that covers broad areas of knowledge and embedded overclaiming in the nomological net of cognitive abilities (*gc* and *gf*), personality (honesty-humility), faking ability, and creative abilities (divergent thinking and originality) and openness-creativity. Next, we discuss our results considering a scaffold of different interpretations of overclaiming behavior and address their implications.

4.1. Overclaiming as a self-enhancement tendency

The first perspective pictures overclaiming as a self-enhancement motivation (Paulhus et al., 2003), that is, describing oneself (deliberately) in a more positive manner. The urge to *fake good* (Ziegler, 2011) and therefore positively misrepresent oneself might be one of the possible explanations for self-enhancement (Bing et al., 2011; Feeney & Goffin, 2015; Kam et al., 2015), as well as the personality factor of honesty-humility. From this perspective, a person claims to have knowledge that they actually do not have (Paulhus, 1984; Paulhus & Reid, 1991) to appear more positive. Hence, we predicted that overclaiming should be more salient in individuals who on the one hand score low in trait honesty-humility as self-serving presentation corresponds with low honesty-humility scores (de Vries et al., 2014), and on the other hand are good at faking (i.e., score high on faking ability). However, our results show that individual differences in overclaiming could neither be explained by honesty-humility scores, nor by the ability to fake. The result for the association between overclaiming and honesty-humility fits into the broad spectrum of research regarding this association so far (e.g., Dunlop et al., 2019; Müller & Moshagen, 2019;

Steger, Schroeders, & Wilhelm, 2020). Although this current null result provokes the endorsement of an interpretation of overclaiming that understands it as more as a mere self-enhancement reflection, we want to mention that our research design was conducted in a low stakes setting, after all.

4.2. Overclaiming as a cognitive bias

Overclaiming has been seen as reflecting some sort of cognitive processing bias (Müller & Moshagen, 2018; Paulhus, 2011), possibly leading to a familiarity experience with terms that is based on morphologic or syntactic characteristics, rather than on actual reflections of one's knowledge (or deliberate faking when reporting one's knowledge). In the present study, the foils used to assess overclaiming were designed to be dissimilar from existing words to prevent remedies between foils and reals from being present in the OCQ—while still being aware that the validity of an overclaiming measure highly depends on the quality of foils (Ackerman & Ellingsen, 2014). Please note that within this approach, it is essential to distinguish between mean effects (i.e., all subjects are equally vulnerable to cognitive biases) and individual differences effects (i.e., measures that differ with respect to provoking cognitive biases are not perfectly related to each other and differ in their correlations with covariates). Obviously, one option to further test this theoretical stance in future studies is to manipulate morphologic and syntactic features of foils. Tricking participants into believing they know certain foils that mimic existing terms raises the question of why some participants are more prone to produce biased answers than others. With our present results, it is difficult to clearly support or refute this notion, but we are confident that future studies will more thoroughly investigate this idea.

4.3. Overclaiming as a proxy for cognitive abilities

In line with the literature, we again found that mean familiarity ratings of foils were unrelated with knowledge (Hülür et al., 2011, 2014). By embedding overclaiming into a nomological net of cognitive abilities, we were able to show again that overclaiming strongly predicts self-reported knowledge (Hülür et al., 2011, 2014), whereas *gc* moderately predicted self-reported knowledge. These findings are in line with previous meta-analytic findings (Freund & Kasten, 2012), implying that the immediate self-report of knowledge in the form of rating one's familiarity with real words is biased and, arguably, of limited utility in an applied context. With the current data, we cannot support the idea that self-reported knowledge in terms of mean familiarity ratings offer an economical tool for indexing cognitive abilities, since self-reported knowledge is confounded by overclaiming and only moderately explained by actual knowledge. By and large, overclaiming and self-reported knowledge are akin to measures of self-reported typical behavior—a measurement approach with which constructs of maximal cognitive effort such as fluid or crystallized intelligence cannot be captured. We suggest that—especially in light of previous studies also reporting low or moderate associations between overclaiming, OC bias, self-reported knowledge, OC accuracy, and *gf* and *gc* (Hülür et al., 2011, 2014; Paulhus & Harms, 2004)—overclaiming as a phenomenon has a more complex nature: if someone is knowledgeable, overclaiming might tell us more about their personality, as this person might overclaim intentionally in order to appear positively. In contrast to that, overclaiming might reflect something more akin to maximal performance if someone less capable authentically reports what he or she think they know. Correspondingly, future studies investigating overclaiming should carefully consider on how to approach the assessment of this phenomenon.

4.4. Overclaiming as a form of creative engagement

Finally, overclaiming behavior was also discussed as the result of intellectual and creative investment traits (Dunlop et al., 2017; Hülür et al., 2011, 2014). This theoretical stance originates from the idea that overclaiming might arise due to more creative persons letting their fantasy flow. According to this perspective, subjectively new and unknown concepts—such as uncommon words, phrases, or neologisms—attract engagement and elaboration from open-minded, creative people, making them more vulnerable to claim familiarity with foils. However, our results show that overclaiming is not substantially predicted by facets of creative ability (divergent thinking and originality). As participants with the ability to fluently produce divergent solutions to a given problem might more easily sense a familiarity to one of those words, they could also be more prone to overclaim. On the other hand, the results showed that overclaiming was significantly predicted by openness-creativity.

Please note that openness moderately predicted overclaiming. This result joins the ranks of the hitherto inconclusive literature on the relation between openness and overclaiming and corroborates the evidence that there indeed is a weak connection between being open towards the unknown and overclaiming one's level of knowledge. Given the only moderate magnitude of this relation, we infer trying to understand overclaiming exclusively in terms of openness is futile, but that an open personality can contribute to overclaiming behavior.

4.5. Conclusion

Taken together, the test development and the presented pattern of results, which are in line with the literature (e.g., Ackerman & Ellingesen, 2014; Bing et al., 2011; Dunlop et al., 2017; Dunlop et al., 2019; Hülür et al., 2011; Paulhus et al., 2003), support that we developed a psychometrically-sound OCQ. In contrast to the prevalent signal detection scoring, we distinguish self-reported knowledge from overclaiming in terms of the mean familiarity ratings of reals and foils, which is a straightforward approach to assess and score overclaiming. Although a substantial amount of variance between overclaiming and self-reported knowledge was shared, according to our results, this shared variance cannot be attributed to a common underlying factor for both types of items (foils vs. reals). In our study, we applied confirmatory factor analysis and discussed four competing perspectives on overclaiming and its nature. Based on our multivariate study, we were able to test competing perspectives on overclaiming. However, the design of our study did not allow for direct testing of overclaiming as a cognitive processing bias, or for discussing overclaiming in light of various application contexts, such as high stakes settings.

In sum, we have organized the state of the literature into four competing theoretical accounts, applied a comprehensive test battery including a variety of covariates, and evaluated structural equation models designed to capture the competing theoretical perspectives. We have shown that a) overclaiming is neither predicted by honesty-humility nor faking ability and therefore reflects something different than mere self-enhancement tendencies; b) carefully constructed foils allow reliable measurement of overclaiming, which is not best taken as indicating a cognitive processing bias; c) overclaiming is not predicted by crystallized intelligence, is highly predictive of self-reported knowledge, and self-reported knowledge is hence not suitable as an index of or a proxy for cognitive ability (more specifically crystallized intelligence); and d) overclaiming is not predicted by divergent thinking and originality, and moderately predicted by self-reported openness for creativity, indicating that overclaiming does not reflect creative ability. In conclusion, we have shed light into the nomological net of overclaiming by showing that it is highly predictive of what we think we know (self-reported knowledge), and by showing that no single theoretical perspective offers an exhaustive explanation of overclaiming behavior.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.intell.2020.101470.

References

R Core Team (2018). *R: A language and environment for statistical computing (Version 3.5.1) [Computer software]*. R Foundation for Statistical Computinghttps://www.R-project.org/.

Ackerman, P. L., & Ellingsen, V. J. (2014). Vocabulary overclaiming — A complete approach: Ability, personality, self-concept correlates, and gender differences. *Intelligence, 46*, 216–227. https://doi.org/10.1016/j.intell.2014.07.003.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*, 997–1013.

Ashton, M. C., & Lee, K. (2001). A theoretical basis for the major dimensions of personality. *European Journal of Personality, 15*, 327–353. https://doi.org/10.1002/per.417.

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*, 340–345. https://doi.org/10.1080/00223890902935878.

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science, 26*, 1295–1303. https://doi.org/10.1177/0956797615588195.

Bensch, D., Maaß, U., Greiff, S., Horstmann, K. T., & Ziegler, M. (2019). The nature of faking: A homogeneous and predictable construct? *Psychological Assessment, 31*, 532–544. https://doi.org/10.1037/pas0000619.

Bensch, D., Paulhus, D. L., Stankov, L., & Ziegler, M. (2017). Teasing apart overclaiming, overconfidence, and socially desirable responding. *Assessment,* 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.

Borman, W., Kubisiak, C., & Schneider, R. (1999). Work styles. In N. Peterson, M. Mumford, W. Borman, R. Jeanneret, & E. Fleishman (Eds.). *An Occupational Information System for the 21st Century: The Development of O*Net* (pp. 213–226). American Psychological Association.

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

Cole, D. A., Perkins, C. E., & Zelkowitz, R. L. (2016). Impact of homogeneous and heterogeneous parceling strategies when latent variables represent multidimensional constructs. *Psychological Methods, 21*, 164–174. https://doi.org/10.1037/met0000047.

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology, 54*, 304–313. https://doi.org/10.1016/j.jmp.2010.01.001.

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.

Dunlop, P. D., Bourdage, J. S., de Vries, R. E., McNeill, I. M., Jorritsma, K., Orchard, M., ... Choe, W.-K. (2019). Liar! Liar! (when stakes are higher): Understanding how the overclaiming technique can be used to measure faking in personnel selection. *Journal of Applied Psychology.*. https://doi.org/10.1037/apl0000463.

Enders, C. K. (2010). *Applied missing data analysis.* Guilford Press.

Feeney, J. R., & Goffin, R. D. (2015). The Overclaiming questionnaire: A good way to measure faking? *Personality and Individual Differences, 82*, 248–252. https://doi.org/10.1016/j.paid.2015.03.038.

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.

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Manual for kit of reference tests for cognitive factors (revised 1963).* Educational Testing Service.

Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin, 138*, 296–321. https://doi.org/10.1037/a0026556.

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*, 2153. https://doi.org/10.3389/fpsyg.2018.02153.

Grosz, M. P., Lösch, T., & Back, M. D. (2017). The narcissism-overclaiming link revisited. *Journal of Research in Personality, 70*, 134–138. https://doi.org/10.1016/j.jrp.2017.05.006.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55. https://doi.org/10.1080/10705519909540118.

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.

Hülür, G., Wilhelm, O., & Schipolowski, S. (2014). Corrigendum to "Prediction of self-reported knowledge with over-claiming, fluid and crystallized intelligence and typical intellectual engagement" [Learning and Individual Differences 21 (6) (2011) 742–746]. *Learning and Individual Differences, 35*, 153–154. https://doi.org/10.1016/j.lindif.2014.07.014.

Jacobs, K. E., & Roodenburg, J. (2014). The development and validation of the self-report measure of cognitive abilities: A multitrait–multimethod study. *Intelligence, 42*, 5–21. https://doi.org/10.1016/j.intell.2013.09.004.

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Test für das Berliner Intelligenz-Strukturmodell (BIS 4) [Berlin–Structure–Intelligence Test (BIS 4)].* Goettingen: Hogrefe.

Kam, C., Risavy, S. D., & Perunovic, W. Q. E. (2015). Using over-claiming technique to probe social desirability ratings of personality items: A validity examination. *Personality and Individual Differences, 74*, 177–181. https://doi.org/10.1016/j.paid.2014.10.017.

Kellen, D., & Klauer, K. C. (2018). Elementary signal detection and threshold theory. In J. T. Wixted (Ed.). *Stevens' handbook of experimental psychology and cognitive neuroscience* (4th ed.). John Wiley & Sons.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*, 486–507. https://doi.org/10.1177/0049124114543236.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121–1134.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 151–173. https://doi.org/10.1207/S15328007SEM0902_1.

Ludeke, S. G., & Makransky, G. (2016). Does the over-claiming questionnaire measure overclaiming? Absent convergent validity in a large community sample. *Psychological Assessment, 28*, 765–774. https://doi.org/10.1037/pas0000211.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of abilities: A review and meta-analysis. *Journal of Applied Psychology, 67*, 280–296.

MacMillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide.* Lawrence Earlbaum Associates Publishers.

McCrae, R. R., & Costa, P. T. (1989). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of Personality and Social Psychology, 56*, 586–595.

Mesmer-Magnus, J., Viswesvaran, C., Deshpande, S., & Joseph, J. (2006). Social desirability: The role of over-claiming, self-esteem, and emotional intelligence. *Psychological Science,* 336–356.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14*, 858–865.

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.

Müller, S., & Moshagen, M. (2018). Overclaiming shares processes with the hindsight bias. *Personality and Individual Differences, 134*, 298–300. https://doi.org/10.1016/j.paid.2018.06.035.

Müller, S., & Moshagen, M. (2019). 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.

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.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609. https://doi.org/10.1037/0022-3514.46.3.598.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.). *The role of constructs in psychological and educational measurement* (pp. 49–69). Inc: Erlbaum Associates.

Paulhus, D. L. (2011). Overclaiming on personality questionnaires. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.). *New perspectives on faking in personality assessement* (pp. 151–164). Inc: Oxford University Press.

Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence, 32*, 297–314. https://doi.org/10.1016/j.intell.2004.02.001.

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.

Paulhus, D. L., & Petrusic, W. M. (2007). *Measuring individual differences with signal detection analysis: A guide to indices based on knowledge ratings.*

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology, 60*, 307–317.

Pennycook, G., & Rand, D. G. (2019). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality.* https://doi.org/10.1111/jopy.12476 jopy.12476.

Phillips, D. L., & Clancy, K. J. (1972). Some effects of "social desirability" in survey studies. *American Journal of Sociology, 77*, 921–940. https://doi.org/10.1086/225231.

Randall, D. M., & Fernandes, M. E. (1991). The social desirability response bias in ethics research. *Journal of Business Ethics, 10*, 805–817.

Raskin, R. N., & Hall, C. S. (1979). A narcissistic personality inventory. *Psychological Reports. 45*. 590-590 https://doi.org/10.2466/pr0.1979.45.2.590.

Raubenheimer, A. S. (1925). An experimental study of some behavior traits of the potentially delinquent boy. *Psychological Monographs, 34*, 1–107.

Rosseel, Y. (2012). Lavaan: An R package for structural equation Moeling. *Journal of Statistical Software, 48*, 1–36.

Schoppe, K. J. (1975). *Verbaler Kreativitäts-test-VKT: Ein Verfahren zur Erfassung verbal-produktiver Kreativitätsmerkmale [verbal creativity test - VKT: A procedure for assessing*

productive creativity features]. (Verlag für Psychologie CJ Hogrefe).
Selker, R., van den Bergh, D., Criss, A. H., & Wagenmakers, E.-J. (2019). Parsimonious estimation of signal detection models from confidence ratings. *Behavior Research Methods, 51*, 1953–1967.
Silvia, P. J., Kaufman, J. C., Reiter-Palmon, R., & Wigert, B. (2011). Cantankerous creativity: Honesty–humility, agreeableness, and the HEXACO structure of creative achievement. *Personality and Individual Differences, 51*, 687–689. https://doi.org/10.1016/j.paid.2011.06.011.
Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., ... Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts, 2*, 68–85. https://doi.org/10.1037/1931-3896.2.2.68.
Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*, 137–149. https://doi.org/10.3758/BF03207704.
Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language, 70*, 36–52. https://doi.org/10.1016/j.jml.2013.09.005.
Steger, D., Schroeders, U., & Wilhelm, O. (2019). On the dimensionality of crystallized intelligence: A smartphone-based assessment. *Intelligence, 72*, 76–85.
Steger, D., Schroeders, U., & Wilhelm, O. (2020). Caught in the act: Predicting cheating in unproctored knowledge assessment. *Assessment,* 1–14. https://doi.org/10.1177/1073191120914970.
Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99*, 181–198.
Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research, 25*, 195–217. https://doi.org/10.1093/ijpor/eds021.
de Vries, R. E., Zettler, I., & Hilbig, B. E. (2014). Rethinking trait conceptions of social desirability scales: Impression management as an expression of honesty-humility. *Assessment, 21*, 286–299. https://doi.org/10.1177/1073191113504619.
Weiss, S., Steger, D., Kaur, Y., Hildebrandt, A., Schroeders, U., & Wilhelm, O. (2020). On the trail of creativity: Dimensionality of divergent thinking and its relation with cognitive abilities and personality. *European Journal of Personality* (In press).
Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*, 956–972. https://doi.org/10.1177/0013164404268674.
Wilhelm, O., Schroeders, U., & Schipolowski, S. (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. Bis 10. Jahrgangsstufe (BEFKI 8-10) [Berlin test for the assessment of fluid and crystallized intelligence for grades 8-10 (BEFKI 8-10)].* Hogrefe.
Williams, K. M., Paulhus, D. L., & Nathanson, C. (2002). The nature of overclaiming: Personality and cognitive factors. *Poster presented at the 110th annual meeting of the American Psychological Association, Chicago, IL*.
Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*, 201–233. https://doi.org/10.1037/xlm0000732.
Ziegler, M. (2011). Applicant faking: A look into the black box. *Psychologist, 49*, 29–36.
Ziegler, M. (2015). "F*** you, I won't do what you told me!" – Response biases as threats to psychological assessment. *European Journal of Psychological Assessment, 31*, 153–158. https://doi.org/10.1027/1015-5759/a000292.
Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The vocabulary and overclaiming test (VOC-T). *Journal of Individual Differences, 34*, 32–40. https://doi.org/10.1027/1614-0001/a000093.

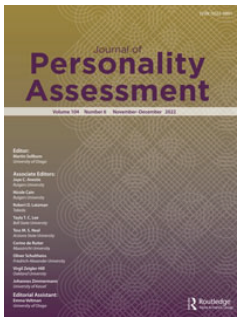
III Manuscript II

The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children

The data for this study has been drawn from an existing data pool of the Hector Research Institute, as provided by co-author Prof. Dr. Jessika Golle, who organized the multivariate study in collaboration with the other co-authors of this manuscript. The overclaiming questionnaire applied in this study was developed by Dr. Stefan Schipolowski, Prof. Dr. Ulrich Schroeders, and Prof. Dr. Oliver Wilhelm. The concept for the study at hand was developed by Prof. Dr. Oliver Wilhelm, Prof. Dr. Ulrich Schroeders, and myself. The data analysis and discussion of results were conducted by myself, mainly under the supervision of Prof. Dr. Oliver Wilhelm and Prof. Dr. Ulrich Schroeders. I have completed a literature search and drafted the original manuscript. All co-authors reviewed and revised the manuscript under my supervision as first author. The manuscript was accepted for publication in the *Journal of Personality Assessment* on the 4th October 2022 and was published online on 28th November 2022.

Republished based on copyright policies from Taylor and Francis; permission conveyed through Copyright Clearance Center, Inc.

Goecke, B., Schroeders, U., Zettler, I., Schipolowski, S., Golle, J., & Wilhelm, O. (2022). The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children. *Journal of Personality Assessment*, 1–12. https://doi.org/10.1080/00223891.2022.2144332



Journal of Personality Assessment

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hjpa20

The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children

Benjamin Goecke, Ulrich Schroeders, Ingo Zettler, Stefan Schipolowski, Jessika Golle & Oliver Wilhelm

To cite this article: Benjamin Goecke, Ulrich Schroeders, Ingo Zettler, Stefan Schipolowski, Jessika Golle & Oliver Wilhelm (2022): The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children, Journal of Personality Assessment, DOI: 10.1080/00223891.2022.2144332

To link to this article: https://doi.org/10.1080/00223891.2022.2144332

View supplementary material

Published online: 28 Nov 2022.

Submit your article to this journal

View related articles

View Crossmark data

This article has been awarded the Centre for Open Science 'Open Data' badge.

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hjpa20

JOURNAL OF PERSONALITY ASSESSMENT
https://doi.org/10.1080/00223891.2022.2144332

Routledge
Taylor & Francis Group

Check for updates

The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children

Benjamin Goecke[1], Ulrich Schroeders[2], Ingo Zettler[3], Stefan Schipolowski[4], Jessika Golle[5], and Oliver Wilhelm[1]

[1]Ulm University, Germany; [2]University of Kassel, Germany; [3]University of Copenhagen, Denmark; [4]Institute for Educational Quality Improvement (IQB), Berlin, Germany; [5]University of Tübingen, Germany

ABSTRACT

Research on self-reported knowledge and overclaiming in children is sparse. With the current study, we aim to close this gap by developing an overclaiming questionnaire measuring self-reported knowledge and overclaiming that is tailored to children. Moreover, we examine the nomological net of self-reported knowledge and overclaiming in childhood discussing three perspectives: Overclaiming as (a) a result of deliberate self-enhancement tendencies, (b) a proxy for declarative knowledge, and (c) an indicator of creative engagement. We juxtaposed overclaiming, as indicated by claiming familiarity with non-existent terms, and self-reported knowledge with fluid and crystallized intelligence, creativity, and personality traits in a sample of 897 children attending third grade. The results of several latent variable analyses were similar to findings known from adult samples: We found no strong evidence for any of the competing perspectives on overclaiming. Just like in adults, individual differences in self-reported knowledge were strongly inflated by overclaiming, and only weakly related to declarative knowledge.

ARTICLE HISTORY
Received 2 January 2022
Accepted 4 October 2022

In many clinical and other both applied and basic research settings, people are asked to self-report their knowledge in certain domains. Exaggerations in measures of self-reported knowledge are also known as overclaiming, a behavioral tendency often measured with overclaiming questionnaires (OCQ; e.g., Paulhus et al., 2003). Overclaiming (OC) is understood as claiming familiarity with terms that do not exist (i.e., foils), whereas self-reported knowledge (SRK) is claiming familiarity with existing terms (i.e., reals). The concept of OC was introduced almost 100 years ago (Raubenheimer, 1925), had a revival in consumer research (Phillips & Clancy, 1972), and has recently received stronger interest, yielding in a surge of measurement instruments (e.g., Atir et al., 2015; Dunlop et al., 2017, 2020; Goecke et al., 2020; Hülür et al., 2011; Müller & Moshagen, 2018b; Ziegler et al., 2013). In contrast to research with adults, experiences with OC and SRK in children are sparse (for an exception, see, e.g., Butler & Nelson, 2021).

Although OCQs were successfully applied in PISA 2012 (a worldwide large-scale assessment of scholastic performances of students in schools; OECD, 2013) with adolescents (e.g., Muszyński, 2020), little is known about this behavior in children. This is somewhat surprising given the fact that the initial study, which laid the foundation for today's OCQs, was conducted in a sample of 10-year-old boys (Raubenheimer, 1925) and was set out to quantify their delinquency. Relatedly, a comprehensive overview of potential explanations for OC in children is still missing, whereas the reasons for OC in adults are discussed more frequently (Dunlop et al., 2020; Goecke et al., 2020; Müller & Moshagen, 2018b). We aim to close this gap and explore the prevalence of OC in children by introducing and psychometrically examining an OCQ for children. Specificallys, we investigate which personality and cognitive ability constructs contribute to individual differences in OC and SRK in children, including openness to experience, honesty-humility, epistemic curiosity, fluid and crystallized intelligence, and divergent thinking.

Scoring of OCQs

Research related to OC is shaped by the use of conventional OCQs, containing both reals and foils (e.g., Paulhus et al., 2003); alternative approaches (e.g., Ackerman & Ellingsen, 2014) are less popular. The prevailing scoring approach for OCQs is based on signal-detection analysis. In this approach, familiarity ratings on reals and foils are translated into dichotomous variables which built the basis for computing *d'*, also called *OC accuracy*, and *c*, also called *OC bias*. These OC technique indices supposedly distinguish a person's sensitivity to accurately categorize reals as existing and foils as non-existing (OC accuracy), and a tendency to claim familiarity with both reals and foils (OC bias; Paulhus, 2011).

CONTACT Benjamin Goecke Benjamin.Goecke@uni-ulm.de Institute for Psychology and Pedagogy, Ulm University, Albert-Einstein-Allee 47, 89081 Ulm, Germany.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/00223891.2022.2144332

© 2022 Taylor & Francis Group, LLC

Both indices were proposed to quantify the tendency to accurately report one's own knowledge and to overclaim in both item types. Although the basic idea might be appealing, utility of both indices for individual differences research has been questioned due to statistical shortcomings (i.e., they are computationally dependent on each other) and their inconclusiveness regarding the underlying cognitive processes (i.e., lower scores in OC bias can be achieved by a better hit rate, by a lower false alarm rate, or both; see Goecke et al., 2020). This notwithstanding, we appreciate the traditional application of this approach in the context of OCQs and thus provide a set of results and a description of the corresponding formulas regarding signal detection indices in the supplementary materials (see SM Figure 1 and SM Table 4).

A more parsimonious and straightforward scoring approach of OCQs uses familiarity ratings on reals and foils as indicators of two experimentally and computationally distinct psychological entities: SRK and OC. SRK is probed by letting participants rate their familiarity with reals and OC is indicated by claiming familiarity with foils (Hülür et al., 2011).

What causes OC behavior in children?

To the best of our knowledge, there is one recent study examining OC in children (Butler & Nelson, 2021). In this study, OC as conceptualized within the signal-detection framework was compared between children and adults by administering a child-friendly OCQ to both groups. Prior to comparing both age groups the convergent validity between the child-friendly OCQ and the OCQ-150 (Paulhus et al., 2003) was investigated in the same sample of adults. Convergent validity, based on correlations between adults' false alarm rates and accuracies (i.e., OC accuracy) only amounted to $r = .51$, and $r = .39$, respectively. Based on the group comparison in the child-friendly OCQ the study concluded that children overclaim more than adults presumably due to different levels of domain-specific knowledge (sensu Ackerman & Ellingsen, 2014). Butler and Nelson (2021) suggest that developmental trajectories in OC should be considered in future research. We suggest caution in generalizing these results, predominantly due to the limited convergent validity of the child-friendly OCQ with the OCQ-150 and the limited evidence concerning invariance of the scale across age.

Several perspectives on what drives OC behavior (in adults) have been proposed. First, OC was interpreted as being driven by self-enhancement (e.g., Paulhus et al., 2003), which might be, among others, related to the personality trait honesty-humility (Goecke et al., 2020). Second, OC accuracy has been described as a proxy for cognitive abilities (Paulhus & Harms, 2004), that is, a useful indicator of a person's knowledge. Third, OC was discussed as some kind of creative engagement (Dunlop et al., 2017) and creative investment traits like openness to experience (Hülür et al., 2011). Although these perspectives are derived by research on OC in adults, they should be a helpful guide for investigating OC in children.

According to the first perspective, people engage in OC because they want to present themselves in a desirable way. One reason for this behavior could be that being (or pretending to be) knowledgeable is socially valued. Connecting OC with a general tendency and proneness to self-enhance is not that simple, however. Theories on socially desirable responding (e.g., Paulhus, 2002) distinguish between intentional and unintentional response behaviors, but typical OCQs cannot evaluate underlying intentions. OC could be seen as driven by self-deceptive (i.e. unintentional) enhancement tendencies (Paulhus & Reid, 1991), but was also proposed to be efficient measuring the systematic (i.e., intentional) tendency to misrepresent oneself (Paulhus et al., 2003; Paulhus & Harms, 2004). The distinction might depend on test settings (e.g., Dunlop et al., 2020). Although the evidence for people showing OC is less convincing in low-stakes than in high-stakes settings (Dunlop et al., 2020), OC seems to be related with narcissism in both (high-stakes: Bensch, Maaß, et al., 2019; low-stakes: Grosz et al., 2017; Ludeke & Makransky, 2016).

Inferring from this view, that is, OC grasps some sort of intentional or unintentional socially desirable responding behavior, OC might be correlated with honesty-humility. This makes sense for two reasons. First, there is a link between aversive personality traits like narcissism and OC, and honesty-humility has been shown to be the opposite pole of a higher-order factor explaining such aversive traits like narcissism (Schreiber & Marcus, 2020). Second, dishonesty explains individual differences in impression management (de Vries et al., 2014), which in turn correlate with OC (Bensch, Paulhus, et al., 2019). Whereas OC supposedly captures exaggeration of what a person knows, low honesty-humility should reflect deceptive or dishonest characteristics more broadly. Although theoretically plausible, the empirical link between OC and honesty-humility is weak in adults (Dunlop et al., 2020; Goecke et al., 2020; Steger et al., 2021). Importantly, though, this association has not been tested in children, so far. If an association between honesty-humility and OC in children was found, this would indicate a change of the association across age and could therefore be taken as an indicator that the reasons for engaging in OC change over the course of development.

Regarding the second perspective, OC and SRK have also been related to fluid (Gf) and crystallized (Gc) intelligence, because SRK and OC accuracy have been proposed as proxies of cognitive abilities (Paulhus & Dubois, 2014; Paulhus & Harms, 2004). It was argued that, because the OC accuracy index is thought to capture a person's sensitivity to accurately categorize reals as existing and foils as non-exisiting, it should be associated with higher levels of knowledge, or (crystallized) cognitive abilities in general (Paulhus & Harms, 2004). This notion remained largely unsupported, however: Correlations between Gc and SRK, or between Gc and OC accuracy usually did not exceed $r = .50$—too little to justify using SRK as a sufficient proxy for Gc. In addition, SRK is inflated by OC, another reason why the hypothesis found little support in the literature (Goecke et al., 2020; Hülür et al., 2011). These incongruencies between self-reported and objectively measured abilities in adults are in line with meta-analyses (Freund & Kasten, 2012) and might also affect children. In fact, children have been reported to be not fully

capable of giving reliable self-estimates about their skills and abilities (Schneider, 2008). Children are overconfident and overestimate their own performances across a wide range of contexts (e.g., cognitive tasks, Finn & Metcalfe, 2014)—effects that supposedly also apply to self-reported knowledge.

Predicting OC and SRK with cognitive abilities (i.e., Gf and Gc) might further our understanding about OC and SRK in children and thus provides an insight into the developmental aspects of OC behavior. For example, the magnitude of the relationship between Gc and SRK could inform us about the amount of congruency between questionnaire and test data of the same underlying latent variable. In this vein, it might also be interesting to study gender differences of SRK and OC. More specifically, a majority of studies using latent variable approaches for depicting Gc indicate a male advantage in children aged 5 to 16 years (Härnqvist, 1997; Keith et al., 2008; Reynolds et al., 2008, 2013). The fact that adolescent boys were also found to be more likely than girls to engage in OC (Vonkova et al., 2018) begs the question whether these mean differences can still be observed in children after differences in Gc have been accounted for.

The third perspective is that OC might be considered a form of creative engagement, because rating familiarity with non-existing terms could be indicative of rather uncritical thinking style toward new things. This might be shown in relations with creativity (Goecke et al., 2020), but also in associations with broader creative investment traits such as epistemic curiosity (EC) or openness to experience (Bensch, Paulhus, et al., 2019; Dunlop et al., 2017, 2020; Goecke et al., 2020). Although creativity, as measured by divergent thinking, is not related to OC in adults (Goecke et al., 2020), this association has never been examined in children. EC was identified as overlapping with openness to experience and intellect (Mussel, 2013), and is conceptualized as the desire to obtain new knowledge either producing positive experiences of intellectual interest or of reducing undesirable conditions of informational deprivation (Piotrowski et al., 2014). EC motivates knowledge acquisition and exploratory behavior (Berlyne, 1954), and therefore individual differences in this trait might also be associated with OC and SRK. Children high in EC interest might be prone to rate being familiar with foils, because they do not realize the nonexistence of items and are generally fascinated by "new" concepts, which could eventually lead to accepting such terms more easily. In other words, OC behavior could partly indicate curiousness. EC deprivation, by contrast, refers to the idea that one is discouraged because one does not fully understand a concept (Litman, 2008). Consequently, children might erroneously claim familiarity with foils, because they are proactively working against that frustration by claiming familiarity with reals and foils to reduce their uncertainty (i.e., deprivation EC might be associated with OC).

In addition to divergent thinking as an ability measure of creative engagement and epistemic curiosity, OC should be correlated with the broader personality trait of openness to experience. People with high levels in this trait may be more prone to overclaim as they are more willing to encounter, engage in, and elaborate on new and unfamiliar ideas—in this vein, OC could be understood as being particularly uncritical toward terms one has never heard of. The empirical results paint an inconclusive picture regarding this association, however. In fact, correlations between openness to experience and OC are moderate at best (e.g., Dunlop et al., 2017).

Purpose of the current study

Research regarding OC and self-reported knowledge in children is still sparse. We know that children engage in OC, but our understanding about factors possibly driving OC is limited. In addition, little is known about the amount of OC in children, if there are differences between boys and girls, or how reliable OC can be measured in children at all. With the exception for a recent study on OC in children (Butler & Nelson, 2021), there are no published OCQs for children. Additionally, studying OC in children might help to further our limited understanding about the construct. If one of the perspectives regarding OC finds a more pronounced relation between OC and one of the juxtaposed constructs in children in comparison to adults, this would indicate developmental trajectories of this behavior which have been neglected so far.and could then be tested more thoroughly. One purpose of our study is thus to replicate previous findings in a new hitherto understudied sample, namely children.

Our research aims are thus as follows: First, we investigate the psychometric properties of a newly developed child-OCQ with 104 items in a large sample of 3rd graders by fitting a confirmatory factor model, which is analogous to previous measurement models for adults (Goecke et al., 2020; Hülür et al., 2011). We expect that SRK and OC are highly related to each other. Second, we investigate gender differences in OC and SRK by means of multigroup confirmatory factor analysis. Adolescent boys have been found to be more likely to engage in OC (Vonkova et al., 2018), which is why we expect to find slightly more OC in boys than in girls, even after differences in knowledge have been accounted for. Third, we juxtapose both OC and SRK with declarative knowledge and fluid intelligence, divergent thinking, honesty-humility, openness to experience, and epistemic curiosity to describe their location within a nomological net. The research objectives and hypotheses were not preregistered.

Method

Participants and procedure

The data for the current study were collected as part of a larger educational program, the *Hector Children's Academy Program* (for details see Golle et al., 2018), from a subsample of $N = 897$ children ($M = 8.59$, $SD = 0.47$ years; 47.4% female) of complete regular school classes. A booklet design with planned missing data was applied for the study, which is why not all scales were administered to the full sample. Each of 71 classes was tested on two subsequent days and there was a total of 17 booklets, which were randomly assigned to the classes. A professional contractor,

independent of the research team, was responsible for the data collection. Students were not rewarded or graded in any way and parents gave informed consent prior to testing. The study was voluntary for all participants.

Measures

Overclaiming questionnaire

OC and SRK were assessed with a new set of 104 items, containing 86 reals and 18 foils, resembling typical real-to-foil ratios of other OCQs (e.g., Goecke et al., 2020; Paulhus et al., 2003). Items were tailored to the typical knowledge of third graders as indicated in school curricula and can be assigned to one of 17 knowledge areas covering three broad knowledge domains (nature and technology, art and culture, politics and society). The OCQ was administered in two pseudo-parallel forms, only differing in the sequence of items. The response scale was fully labeled and similar to formats that have been successfully used in other studies (e.g., Goecke et al., 2020; Hülür et al., 2011). Participating children were instructed to indicate their familiarity to all presented terms on a 5-point scale (1 = *Never heard of it before*, 2 = *Heard of it before, but don't know anything about it*, 3 = *I know a little about it*, 4 = *I know a fair amount about this*, and 5 = *I know all about this*) within 18 minutes. The presence of foils was not instructed. The OCQ was completed by $N = 897$ children; each item was on average answered by 97.8% of the sample. We present the complete German questionnaire and descriptive statistics for all items in the supplementary materials (SM Table 1).

Fluid and crystallized intelligence

Students' fluid and crystallized intelligence was measured with an age-adapted version of the *BEFKI* (Berlin Test of Fluid and Crystallized Intelligence; see e.g., Schroeders et al., 2020), suited for primary school children (Schroeders et al., 2016). The Gf test was comprised of nine verbal, nine numerical, and 16 figural reasoning items and was administered in two parallel forms. Gc was measured with two parallel test forms compiled of 42 knowledge items per version across three broad content domains (humanities, social sciences, and natural sciences). Gf and Gc were completed by $N = 848$ participants.

Divergent thinking

A verbal divergent thinking test (Carroll, 1993) based on Guilford's *Alternative Uses Task* (Guilford, 1967) was used. The test consisted of two items requiring students to name within 2 min as many ideas as possible about what could be done with a wooden board and a brick, respectively. Answers were counted by three raters after being checked for meaningfulness and duplicates, thus tapping the fluency aspect of divergent thinking (Weiss et al., 2020). $N = 205$ children completed this task.

Personality

Personality was measured as other-report, completed by one or both of the children's parents. We administered a newly developed and age-adapted other report, based on the HEXACO Model of Personality (Ashton & Lee, 2007; Zettler et al., 2020) −60, covering the personality factors of Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience with 10 items each. Parents indicated their agreement with the corresponding statements on a five-point scale ranging from "strongly disagree" to "strongly agree." The other report was filled out for $N = 562$ children by one parent, and sometimes by both parents per child. In the latter case, parents agreed upon their responses before answering the questionnaire.

Epistemic curiosity

We assessed epistemic curiosity with a newly developed and age-appropriate scale containing 14 items. In line with scales measuring epistemic curiosity in adults, each item was supposed to tap either interest or deprivation (Litman, 2008; Litman & Mussel, 2013), resulting in 7 items per dimension. Whereas interest refers to curiosity that is associated with joy regarding new information, deprivation refers to the idea that one is more concerned with decreasing precariousness. Items were statements (e.g., "*I am pleased to learn something new.*") and children had to indicate how often these statements applied to them on a 4-point scale (1 = *never*, 2 = *seldom*, 3 = *often*, 4 = *always*). The epistemic curiosity questionnaire was completed by $N = 192$ children.

Statistical analyses

All analyses were conducted with R (R Core Team, 2020) in the framework of *Structural equation modeling* (SEM) with the R package *lavaan* (Rosseel, 2012). To reduce model complexity, we used item parcels as indicators after establishing the unidimensionality of the scales (Little et al., 2002). More detailed descriptions of the parcels, the unidimensional measurement models, reliabilities per scale, as well as information regarding the preparation of the data and the computation of parcels can be found in the supplementary materials. We modeled OC and SRK as latent variables with parcels of foils and reals, respectively, as indicators. The six parcels per scale were built randomly, but in a way that each parcel was stratified across all broad content domains (Cole et al., 2016). Based on these parcels, we tested for possible effects of the nested data structure (children nested in classes) but found on average only 2.7% of variance in these indicators explained by the class-level, whereas the rest of the variance was accounted for by the individual-level. We thus ignored clustering in the subsequent analyses (Pornprasertmanit et al., 2014). Where possible, we used *full information maximum likelihood* estimation under the assumption that missing information was missing completely at random to combine missing data handling and parameter estimation in a single step (Schafer & Graham, 2002). Models based on dichotomous indicators are based on the *Weighted Least Squares*

Table 1. Descriptive statistics for familiarity ratings, including the means, standard deviations, skewness, and kurtosis for each OCQ parcel.

			Reals			Foils		
Parcel	*n* (reals)	*n* (foils)	*M (SD)*	Skewness	Kurtosis	*M (SD)*	Skewness	Kurtosis
1	15	3	2.97 (.85)	.33	−.46	2.00 (1.16)	1.10	0.20
2	15	3	3.00 (.81)	.30	−.29	1.76 (1.03)	1.47	1.33
3	14	3	2.96 (.89)	.29	−.69	1.79 (1.09)	1.50	1.39
4	14	3	2.49 (.84)	.92	.47	1.67 (1.02)	1.72	2.23
5	14	3	3.18 (.80)	.07	−.42	1.74 (1.05)	1.62	1.88
6	14	3	2.76 (.87)	.59	−.16	1.78 (1.05)	1.44	1.26

Table 2. Model fit for the measurement models of all scales.

Measurement model		*n* (Parcels)	*n* (Persons)	χ^2	*df*	CFI	RMSEA	[90% CI]	SRMR	ω	*r* (*SE*)
1	OC	6	897	18.20	9	.993	.034	[.017; .050]	.010	.94	
2	SRK	6	897	55.51	9	.992	.076	[.059; .094]	.009	.97	
3	Gf/Gc	6	848	28.63	8	.984	.055	[.034, .078]	.023	.69/.74	.77 (.03)
4	EC Int/Dep	6	210	13.31	8	.986	.056	[.000; .103]	.032	.82/.73	.82 (.05)
5	Honesty-Humility	4	578	0.47	2	1.00	.000	[.000; .051]	.013	.81	
6	Openness to Experience	4	577	5.39	2	.990	.054	[.000; .107]	.017	.73	

Note. EC = Epistemic Curiosity; Int = Interest; Dep = Deprivation. For both Models 3 and 4, two correlated factors were specified. ω = McDonald's Omega (McDonald, 1999).

Means and Variance adjusted (WLSMV) estimator (Beauducel & Herzberg, 2006); in this case, pairwise observations were used. Models with continuous indicators are based on a maximum likelihood estimator with robust standard errors (MLR). Good model fit was indicated by CFI (*Comparative Fit Index*) $\geq$ .95, RMSEA (*Root Mean Square Error of Approximation*) $\leq$ .06, and the SRMR (*Standardized Root Mean Square Residual*) $\leq$ .08 (Hu & Bentler, 1999).

Gender differences in OC and SRK when controlling for Gc levels were studied in a multigroup confirmatory factor analysis (MGCFA) using a reference group approach with boys being the reference group (Schroeders & Gnambs, 2020). We tested necessary measurement invariance conditions with progressively restrictive stages (Vandenberg & Lance, 2000). First, we conducted a configural invariance test to establish a baseline model across groups. Second, we tested for metric invariance and constrained factor loadings to equality across groups. Third, we conducted a scalar invariance model, in which the factor loadings and the intercepts were constrained to equality across groups. This type of invariance is the precondition for comparing latent group means. The standardized latent mean of the comparison group can be interpreted as Cohen's *d* (Cohen, 1988), given that the reference group identification approach has been used. Lastly, we checked for strict measurement invariance, by adding restrictions on the residuals across the groups. Invariance tests were examined by assessing changes in the CFI with values >.01 (Chen, 2007) as indication of substantial model deterioration.

To make all analyses reproducible, we provide all materials necessary, including a complete correlation matrix with descriptive statistics for all variables, and data along with the syntax for the statistical analysis in an online repository: https://osf.io/94bk8/

Results

We report means and standard deviations of mean familiarity ratings of reals and foils comprising the broad knowledge domains of the OCQ (Table 1). Descriptive statistics are comparable to those found in adult samples, that is, the range of the statistics of the parcels is similar to values of other OCQs used in the literature (Dunlop et al., 2020; Goecke et al., 2020; Hülür et al., 2011). Please note that the kurtosis of the foils was lighter tailed in our study, however, indicating somewhat more positive familiarity ratings (OC) of the sample.

Measurement models

Fit indices of unidimensional measurement models based on item parcels for all measures (see Table 2) indicate good model fit, with the exception of the RMSEA for the SRK model (Model 2). The reliability of all latent variables was acceptable. Measurement models for intelligence (Gf and Gc) and epistemic curiosity included a correlation between the two postulated latent factors, which were in the expected range (e.g., Litman, 2008; Schroeders et al., 2016). We do not report fit indices for the measurement model of fluency, because this scale only contained two indicators.

A latent regression model, in which SRK is predicted by OC (see Figure 1), was previously established in adult samples (Goecke et al., 2020; Hülür et al., 2011) and is the basis for all following models. OC predicting SRK considers that an individual's SRK might be affected by OC. OC accounted for 62.3% of the variance in SRK. In comparison, this regression was somewhat weaker in two adult samples where OC accounted for 50.5% (Goecke et al., 2020) and 43.6% of SRK (Hülür et al., 2011), respectively. This indicates that SRK in children might be more inflated by OC in comparison to adults.

Juxtaposing OC and SRK with the parent-rated personality traits honesty-humility and openness to experience represent the perspective of OC being the result of an intentional strategy to manage others impressions or an open-minded personality (Figure 2). However, OC was neither predicted by openness to experience nor by honesty-humility, whereas SRK was significantly but only infirmly predicted by

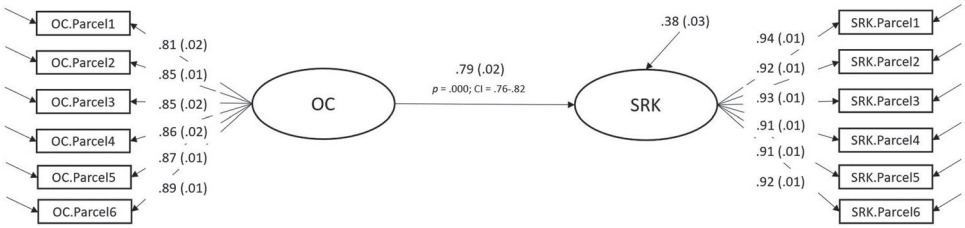


Figure 1. Structural model of OC and SRK.
Note. CI = 95% Confidence Interval. All parameters are standardized. $n = 897$; $\chi^2(53) = 396.1$, CFI = .962, RMSEA = .085, SRMR = .040.

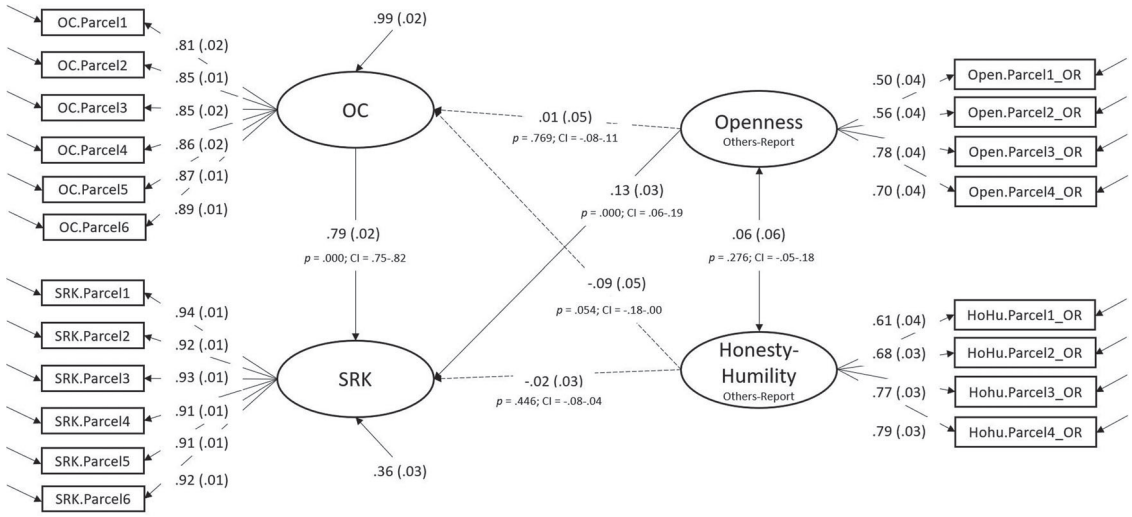


Figure 2. Structural model of OC, SRK, openness to experience, and honesty-humility.
Note. CI = 95% Confidence Interval. All parameters are standardized. Non-significant coefficients are displayed with dashed lines ($p > .05$). $n = 897$; $\chi^2(164) = 626.8$, CFI = .960, RMSEA = .056, SRMR = .040.

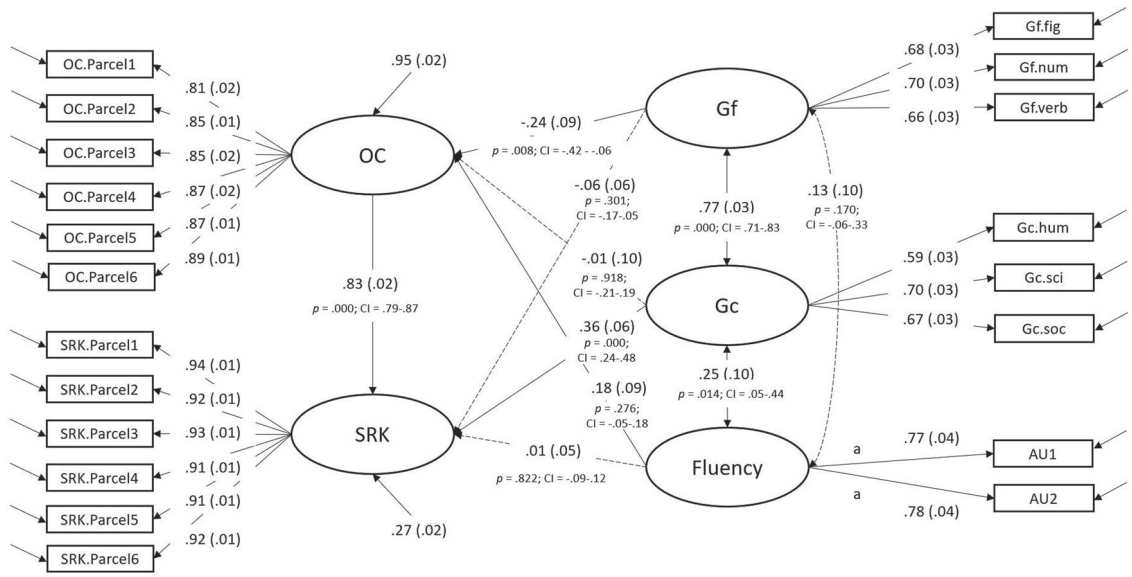


Figure 3. Structural model of OC, SRK, Gc, Gf, and fluency.
Note. CI = 95% Confidence Interval. All parameters are standardized. Non-significant coefficients are displayed with dashed lines ($p > .05$). $n = 897$; $\chi^2(161) = 652.7$, CFI = .961, RMSEA = .058, SRMR = .041.

openness to experience. The non-substantial regression coefficients from honesty-humility and openness to experience to OC is in line with earlier findings with adults (e.g., Dunlop et al. 2017; Goecke et al., 2020), the non-substantial regression weight from openness to experience toward OC adds evidence to earlier, inconsistent findings.

Next, we juxtaposed the previously established measurement models of OC and SRK with Gf, Gc, and fluency (see Figure 3). Fluency was modeled together with Gf and Gc, because all three constructs should be understood as abilities that require maximal performance behavior. Only 8% of the variance in OC was explained, with small regression coefficients of Gf and fluency toward OC. Gf predicted OC

negatively and fluency predicted it positively. Gc did not account for individual differences in OC. These findings slightly deviate from findings in adults, where OC was unrelated to all three constructs (Goecke et al., 2020; Hülür et al., 2011).

SRK was more strongly explained by OC than by actual knowledge. In total, 72.3% of the variance in SRK was explained by OC and Gc, again resembling findings for adult samples (Goecke et al., 2020; Hülür et al., 2011). The moderate prediction of SRK by Gc is in line with meta-analytical findings regarding the congruency of self-reported and objectively measured abilities (e.g., Freund & Kasten, 2012).

Latent mean differences between boys ($n = 447$) and girls ($n = 423$) in OC and SRK were studied by first fitting a slightly adjusted model in which OC and SRK were predicted by Gc, and in which the residuals of both latent factors were correlated. Because a strict measurement invariant model holds (see SM Table 3 in the supplementary materials), means of all factors can be interpreted. As we used boys as the reference group, the latent mean parameter estimates of the girls depict the gender differences (i.e., standardized mean differences; see parameter estimates in Table 3). Looking at the results, girls had lower means on all factors with small effect sizes (see the *d* estimates in Table 3). In line with the literature, boys showed slightly more actual knowledge (Gc) than girls (Keith et al., 2008). However, boys also tend to overclaim more knowledge than girls, as depicted in higher scores on OC and SRK (see also Vonkova et al., 2018). The relations between the three postulated factors were by and large very similar across groups, indicating that the self-reported knowledge scores of girls are as inflated by overclaiming as the boys' scores, even when controlling for Gc.

Finally, OC and SRK were regressed onto epistemic curiosity (see Figure 4). Both factors of epistemic curiosity did not account for substantial shares of variance in OC and SRK.

Table 3. Parameter estimates of strict invariance model depicting OC, SRK, and Gc.

	Standardized mean differences (*d*)	95%-CI		
SRK	−.26	[−.41; −.10]		
OC	−.28	[−.45; −.11]		
Gc	−.21	[−.37; −.05]		
	Standardized estimates			
	Boys	95%-CI	Girls	95%-CI
r (SRK, OC)	.86	[.82; .90]	.82	[.78; .86]
β (OC ~ Gc)	−.17	[−.28; −.06]	−.19	[−.30; −.07]
β (SRK ~ Gc)	.21	[.10; .32]	.15	[.03; .26]

Note. *d* = Cohen's *d*. The standardized mean differences depict the standardized means of the girls (boys as reference). Negative values of *d* indicate lower means for girls, whereas positive values of *d* indicate higher means for girls.

Discussion

The literature concerning OC and SRK in adults is broad and so are the corresponding interpretations of this behavior. OC has been mostly studied in normal populations, but is has also been linked to possibly maladaptive personality traits like narcissism (e.g., Grosz et al., 2017) or problematic self-enhancement tendencies (e.g., Paulhus et al., 2003) that might distort self-report questionnaires. Surprisingly, little is known about this behavior in children, even though the construct originated in child development research, where it was introduced as a measure of deviant behavior of boys (Raubenheimer, 1925).

The aims of the present study were (a) to develop an OCQ for children and to evaluate its psychometric properties, and (b) to study gender differences in OC and SRK, and place both constructs into a broad nomological net. We thus developed an OCQ for children tapping 17 specific knowledge areas from three broad knowledge domains. We embedded both OC and SRK in the nomological net of maximal cognitive performance (Gc, Gf, and fluency) and typical behavior (honesty-humility, openness to experience, and epistemic curiosity). In the next sections, we summarize and discuss our findings and sketch their implications.

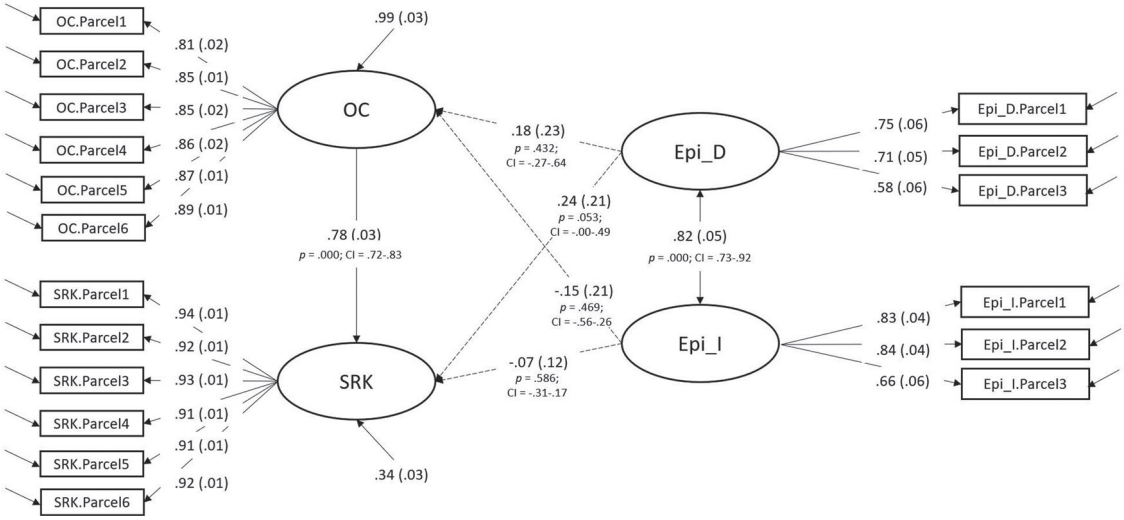


Figure 4. Structural model of OC, SRK, and epistemic curiosity.
Note. Epi_D = Epistemic Curiosity Deprivation; Epi_I = Epistemic Curiosity Interest; CI = 95% Confidence Interval. Parameters are standardized; non-significant coefficients are displayed with dashed lines ($p > .05$). $n = 897$; $\chi^2(153) = 559.6$, CFI = .961, RMSEA = .061, SRMR = .042.

Perspectives on overclaiming

The pattern of results support that we developed a psychometrically sound OCQ. We used reals and foils to distinguish between OC and SRK. This is a straightforward approach to score an OCQ. The range of the descriptive statistics for all items and parcels were comparable to those of OCQs administered in adult samples (e.g., Goecke et al., 2020; Hülür et al., 2011). The newly developed child-OCQ exhibits the same structure as OCQs for adults (e.g., Goecke et al., 2020; Hülür et al., 2011), namely, that responses on the OCQ can be modeled with two factors, where one factor captures the communality of the foils and the other factor captures the communality of the reals. Moreover, SRK of children was considerably affected by OC, with slightly more variance explained in SRK than usually found in adults. This should be taken as evidence that children already show OC (see also Butler & Nelson, 2021) and confirms studies showing that children tend to overestimate their abilities (e.g., Finn & Metcalfe, 2014; Schneider, 2008). In summary, it is possible to measure OC and SRK with a conventional OCQ in children.We tested competing perspectives on why persons engage in OC (Goecke et al., 2020) by putting the most prevalent interpretations of OC in adults to a test in children. The overall aim was to lay the foundation for a better understanding of OC by relating the results found in adults to the ones found in children.

A first perspective predicts that OC should be related to honesty-humility, if persons intentionally claim knowledge they cannot have (Paulhus, 1984). Comparable to results from OC in adults (e.g, Dunlop et al., 2020; Goecke et al., 2020; Müller & Moshagen, 2018a; Steger et al., 2021), this was not the case in our study with children. OC was unrelated with honesty-humility, suggesting that dishonest children typically do not engage more or less in OC than others. Although small but substantial relations between OC and more aversive personality traits like narcissism exist, the lack of an association between OC and honesty-humility suggests that OC occurs unintentionally in children. It has to be noted, however, that honesty-humility rather reflects preferences for interpersonal behavior and less intentions to present oneself in a better light, which is why the perspective of OC being reflective of intentional tendencies to misrepresent oneself should not be disregarded in the future.

We also found no support for the second perspective, which connects the responses on OCQs with cognitive abilities—particularly Gc (Paulhus & Dubois, 2014; Paulhus & Harms, 2004). OC was unrelated to Gc and only weakly predicted by Gf. SRK was more inflated by OC than it was predicted by Gc. The magnitude of the relation between SRK and Gc underlines that SRK cannot serve as a valid proxy for actual knowledge. Gf and fluency were essentially unrelated with SRK. Thus, we argue that self-reports of cognitive abilities in children cannot serve as a valid measure of these constructs, similar to adult samples although some overlap between these constructs exist (e.g., Freund & Kasten, 2012). SRK is better understood as a misjudgment and misperception regarding familiarity with existing terms instead. Gender differences in OC and SRK after controlling for individual differences in Gc showed that boys tended to overclaim more and also indicated more SRK (e.g., Vonkova et al., 2018). When considering the Gc levels, however, it becomes apparent that the amount of variance in SRK explained by OC is similar for both genders. Thus, both girls and boys overestimate their knowledge when rating familiarity with reals.

Lastly, we examined the perspective that OC is associated with forms of creative engagement and creative investment traits such as openness to experience or EC (e.g., Dunlop et al., 2017; Goecke et al., 2020; Hülür et al., 2011). This perspective takes OC as indicative of an open-minded manner to be familiar with unfamiliar terms (i.e., predominantly foils), referring to the idea that new but yet unknown terms (such as foils) might be "enriched" by subjects scoring higher in openness to experience, epistemic curiosity, or creativity, due to their inherent willingness to find meaning in new things. Importantly, though, this tendency could also be seen as an uncritical adoption of concepts that do not exist. We tested this idea by regressing both OC and SRK onto openness to experience and epistemic curiosity on the personality side, and onto fluency on the maximal cognitive performance side. By and large, our results add to the conclusion that this perspective lacks support: Openness to experience was not predictive of OC, and only weakly predicted SRK. Both facets of epistemic curiosity, interest and deprivation, explained no variance in OC and SRK. In contrast to an adult sample (Goecke et al., 2020), fluency as an indicator of creativity was not predictive of SRK, but was able to explain a small amount of variance in OC. In total, we found no support for this interpretation of OC.

Apart from the interpretations of OC taken above, we argue that at least some mechanisms underlying OC in children should be attributed to their developing metacognition (Schneider, 2008). Children with highly developed metacognitive knowledge can be considered experts regarding their own knowledge—they know if they know something or not—which is readily related to OC. In the same vein, it was argued that overclaiming should be connected to metacognitive knowledge (Clariana et al., 2016), because OCQs tap into the so-called *feeling of knowing* (Paulhus, 2011), which is considered a part of metacognitive knowledge. In theory, this is the part of metacognitive knowledge that helps people decide if they know or do not know a specific information. This is exactly what is required of participants when they have to fill out an OCQ. They see terms and have to decide whether they are familiar with them or not. In children, declarative metacognitive knowledge increases with age and is associated with age-related improvements in cognitive and school performances (Schneider & Löffler, 2016). In other words, the developments of cognitive abilities and metacognitive knowledge seem to be intertwined. Thus, future studies examining OC in children should consider a metacognitive perspective on OC.

On the Use of OCQs

In the present study, we showed that overclaiming expressed as rating the familiarity with foils can be reliably measured

in children and is strongly related with self-reported knowledge. With respect to the validity of OC, we replicated earlier critical results established in adult samples. The nomological net of OC, as we tested it, provides insufficient support for the different perspectives on OC that have been proposed in the literature. In turn, one might argue that one should not only conceptualize OC as measured by conventional OCQs as a valuable construct of individual differences since the evidence so far regarding its predictive validity for real-world outcomes is rather stuttering (Watts et al., 2021), although some studies give reason to be optimistic (e.g., Bing et al., 2011; Dunlop et al., 2017). In order to make sure that this conclusion is not premature, studies that systematically vary test design or juxtapose the OCQ with other measures of OC might be warranted. In the following, we present a number of related research questions that are to our knowledge largely unanswered.

Research related to OC is shaped by the use of conventional setups of OCQs, that is, it is characterized by a number of default settings that might or might not be optimal in a psychometric sense. This is reflected in a large number of different OCQs described in the literature (e.g., Atir et al., 2015; Butler & Nelson, 2021; Dunlop et al., 2017, 2020; Goecke et al., 2020; Hülür et al., 2011; Mesmer-Magnus et al., 2006; Paulhus et al., 2003; Ziegler et al., 2013). These OCQs usually comprise 80% reals and 20% foils (Paulhus, 2011), and items are restricted to a number of knowledge domains (e.g., Paulhus et al., 2003). The fixed ratio of reals and foils might be seen as an unnecessary restriction that rests on the presupposition that other ratios will not change the rank order of participants. Other things being equal, a more conservative approach of expressing familiarity will presumably be beneficial if the base rate of foils is high. Moreover, high conceptual proximity, relative plausibility, or only orthographical resemblances of foils to existing concepts is likely to deceive otherwise knowledgeable participants into endorsing familiarity with foils. The question how foils should be constituted is not conclusively answered. In addition to that, most OCQs are administered without informing participants about the presence of foils and their base rate (e.g., Paulhus et al., 2003). Knowledgeable and intelligent participants might infer the presence of foils, specifically in domains in which they have expertise. This inference partly removes blindness of participants toward their treatment and participants that infer the presence of foils are likely to adjust their response behavior. Taken together, systematically varying the base rate of foils and manipulating the information about their presence in the instructions should help us understand to what extent results found in studies with OCQs are generalizable.

More general, the method of OCQs should be scrutinized critically. The basic idea of rating familiarities with reals and foils is hardly different from conventional vocabulary tests, although conventional vocabulary tests show considerably more convergent validity with other measures of Gc (Schipolowski et al., 2014). The traditional OCQ format could be varied to gradually shift OCQs toward vocabulary tests. For instance, the response formats of both tests could be aligned, in order to test the relation of OCQs to more traditional vocabulary tests depending on different response formats: word/non-word decisions, finding synonyms or antonyms, choosing the existing word among non-words, and finally: decide whether one knows a term or not.

For the most part, OCQs have been applied in the context of traditional knowledge domains, often by inquiring about the familiarity with single words. The rational of the technique also allows studying familiarity with factitious or fictitious statements and measures could also include less traditional domains such as familiarity with dance moves, novel music, rumors about celebrities etc. So far, we have limited knowledge whether or not the technique is more useful in these less academic contexts.

There are also a number of methodological challenges of the prevalent signal detection approach (Paulhus, 2011) that have not been addressed adequately yet. For example, there is no conventional measurement model. The coherence of both reals and foils is neglected. The relation between reals and foils and how and why this relation might vary on an item level is also neglected although this covariance is clearly essential for the traditional indices derived in OCQs (i.e., the hit rate, the false-alarm rate, and eventually *d'* and *c*). These indices, which are ratio variables, are logically and computationally dependent (Zimprich et al., 2020). These issues deserve more attention if the signal detection approach is meant to continue to be the method of choice for scoring OCQs.

In summary, the utility of the OCQ-technique strikes us as limited so far. The suggestions presented in this section might help to endorse a more final verdict, one way or another.

Where do we go from here?

The previous literature on OC seems to be consensual about the fact that we can reliably observe individual differences in OC as measured with OCQs across numerous contexts (i.e., low- and high-stakes settings). The construct of OC has gained considerable attention in the last 10 years—not to mention its' application in PISA 2012. Unfortunately, however, individual differences in OC are not stably associated with other constructs of interest—both personality traits and cognitive abilities—across various studies. This raises the question of what OC actually reflects. Empirically supported answers to this question should be conceived as a presupposition for using OC in applied contexts. Large multivariate studies measuring OC might help us to better understand the very nature of this behavior, but we argue that these studies should not be limited to a single operationalization of OC.

In fact, research on a construct should usually transcend the methods of inquiry. OC should go beyond what a test presumably measures—i.e. exaggerating one's knowledge by falsely claiming familiarity with a word. Other operationalizations of OC are thinkable and desirable. If the purpose of a measurement instrument is to detect exaggerations regarding a person's knowledge, then OC could also be measured, for example, by letting participants first answer knowledge questions by stating whether or not they know the correct response and have them at a later stage answer the questions

by retrieving or recognizing the correct response. Another method could be to ask participants to respond to fictional knowledge questions in an open-ended response format, for example, "Which researcher won the Nobel prize for physics in 1700?." Participants who do not stat "I don't know" or "There is no answer to this question" are then engaging in OC. Relatedly, the combination between actual knowledge tests and self-ratings regarding one's own knowledge could also initiate research toward a possible counterpart of OC; underclaiming, or belittling oneself. It seems fruitful to vary the response format of instruments measuring OC to further the understanding of the construct. Thus, it would be possible to disentangle the measurement from the construct. From an overarching perspective, engaging in self-serving and exaggerating behavior also comprises tweaking biographical events, malingering symptoms for gaining monetary benefits, presenting character strengths in speed-dating, and the like. The broadness of this domain is not reflected in available measurement instruments for OC, however, although OC might be relevant behavior in these contexts. In light of our current results and the previous literature on OC, we thus argue that a multi-method approach should be warranted for furthering our understanding about overclaiming as a construct of individual differences. Such an approach might also inform us about the utility of measuring OC for diagnostic purposes in clinical and other applied settings, like high-stake application contexts where identifying distorted self-reports and individuals engaging in faking might be helpful.

Conclusion

The present study contributes to the hitherto sparse literature on OC and SRK in children. Children seem to overclaim their knowledge at least as much as adults do, but the predictors of this behavior are unknown in both children and adults. OC should be considered a reliable latent variable of individual differences, but it has little to do with honesty-humility, cognitive abilities, or creative investment traits. If OC will be studied in the future, it should be investigated with a more systematic variation of the stimulus material and with multiple methods—in both children and adults.

Disclosure statement

The authors have no conflicts of interest to disclose.

Funding

The work reported herein was supported by grants from the Hector Foundation II.

Open Scholarship



This article has earned the Center for Open Science badges for Open Data through Open Practices Disclosure. The data are openly accessible at https://osf.io/94bk8/.

ORCID

Benjamin Goecke http://orcid.org/0000-0002-3050-1848
Ulrich Schroeders http://orcid.org/0000-0002-5225-1122
Ingo Zettler http://orcid.org/0000-0001-6140-7160
Jessika Golle http://orcid.org/0000-0002-8507-7079
Oliver Wilhelm http://orcid.org/0000-0001-7980-1166

Data accessibility statement

The study materials, data, and analysis scripts used for this article can be accessed at https://osf.io/94bk8/

References

Ackerman, P. L., & Ellingsen, V. J. (2014). Vocabulary overclaiming—A complete approach: Ability, personality, self-concept correlates, and gender differences. *Intelligence*, *46*, 216–227. https://doi.org/10.1016/j.intell.2014.07.003

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

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, *26*(8), 1295–1303. https://doi.org/10.1177/0956797615588195

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Bensch, D., Maaß, U., Greiff, S., Horstmann, K. T., & Ziegler, M. (2019). The nature of faking: A homogeneous and predictable construct? *Psychological Assessment*, *31*(4), 532–544. https://doi.org/10.1037/pas0000619

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

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology*, *43*(3), 180–191.

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

Butler, J. R., & Nelson, N. L. (2021). Children overclaim more knowledge than adults do, but for different reasons. *Journal of Experimental Child Psychology*, *201*, 104969. https://doi.org/10.1016/j.jecp.2020.104969

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Clariana, M., Castelló, A., & Cladellas, R. (2016). Feeling of knowing and over-claiming in students from secondary school to university. *Learning and Individual Differences*, *49*, 421–427. https://doi.org/10.1016/j.lindif.2016.05.008

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Cole, D. A., Perkins, C. E., & Zelkowitz, R. L. (2016). Impact of homogeneous and heterogeneous parceling strategies when latent variables represent multidimensional constructs. *Psychological Methods*, *21*(2), 164–174. https://doi.org/10.1037/met0000047

de Vries, R. E., Zettler, I., & Hilbig, B. E. (2014). Rethinking trait conceptions of social desirability scales: Impression management as an

expression of honesty-humility. *Assessment*, *21*(3), 286–299. https://doi.org/10.1177/1073191113504619

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

Dunlop, P. D., Bourdage, J. S., de Vries, R. E., McNeill, I. M., Jorritsma, K., Orchard, M., Austen, T., Baines, T., & Choe, W.-K. (2020). Liar! Liar! (when stakes are higher): Understanding how the overclaiming technique can be used to measure faking in personnel selection. *The Journal of Applied Psychology*, *105*(8), 784–799. https://doi.org/10.1037/apl0000463

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning and Instruction*, *32*, 1–9. https://doi.org/10.1016/j.learninstruc.2014.01.001

Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, *138*(2), 296–321. https://doi.org/10.1037/a0026556

Goecke, B., Weiss, S., Steger, D., Schroeders, U., & Wilhelm, O. (2020). Testing competing claims about overclaiming. *Intelligence*, *81*, 101470. https://doi.org/10.1016/j.intell.2020.101470

Golle, J., Zettler, I., Rose, N., Trautwein, U., Hasselhorn, M., & Nagengast, B. (2018). Effectiveness of a "Grass Roots" statewide enrichment program for gifted elementary school children. *Journal of Research on Educational Effectiveness*, *11*(3), 375–408. https://doi.org/10.1080/19345747.2017.1402396

Grosz, M. P., Lösch, T., & Back, M. D. (2017). The narcissism-overclaiming link revisited. *Journal of Research in Personality*, *70*, 134–138. https://doi.org/10.1016/j.jrp.2017.05.006

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill, Inc.

Härnqvist, K. (1997). Gender and grade differences in latent ability variables. *Scandinavian Journal of Psychology*, *38*(1), 55–62. https://doi.org/10.1111/1467-9450.00009

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

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

Keith, T. Z., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock–Johnson III tests of cognitive abilities. *Intelligence*, *36*(6), 502–525. https://doi.org/10.1016/j.intell.2007.11.001

Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences*, *44*(7), 1585–1595. https://doi.org/10.1016/j.paid.2008.01.014

Litman, J. A., & Mussel, P. (2013). Validity of the interest-and deprivation-type epistemic curiosity model in Germany. *Journal of Individual Differences*, *34*(2), 59–68. https://doi.org/10.1027/1614-0001/a000100

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 151–173. https://doi.org/10.1207/S15328007SEM0902_1

Ludeke, S. G., & Makransky, G. (2016). Does the Over-Claiming Questionnaire measure overclaiming? Absent convergent validity in a large community sample. *Psychological Assessment*, *28*(6), 765–774. https://doi.org/10.1037/pas0000211

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

Mesmer-Magnus, J., Viswesvaran, C., Deshpande, S., & Joseph, J. (2006). Social desirability: The role of over-claiming, self-esteem, and emotional intelligence. *Psychological Science*, *48*(3), 336–356.

Müller, S., & Moshagen, M. (2018a). True virtue, self-presentation, or both?: A behavioral test of impression management and overclaiming. *Psychological Assessment*, *31*(2), 181–191. https://doi.org/10.1037/pas0000657

Müller, S., & Moshagen, M. (2018b). Overclaiming shares processes with the hindsight bias. *Personality and Individual Differences*, *134*, 298–300. https://doi.org/10.1016/j.paid.2018.06.035

Mussel, P. (2013). Intellect: A theoretical framework for personality traits related to intellectual achievements. *Journal of Personality and Social Psychology*, *104*(5), 885–906. https://doi.org/10.1037/a0031918

Muszyński, M. (2020). *Validity of the overclaiming technique as a method to account for response bias in self-assessment questions*. Analysis on the basis of the PISA 2012 data [University at Krakow]. https://ruj.uj.edu.pl/xmlui/bitstream/handle/item/251494/muszynski_validity_of_the_overclaiming_technique_as_a_method_2020.pdf?sequence=1&isAllowed=y

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD. https://doi.org/10.1787/9789264190511-en

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Erlbaum Associates, Inc.

Paulhus, D. L. (2011). Overclaiming on personality questionnaires. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessement* (pp. 151–164). Oxford University Press, Inc.

Paulhus, D. L., & Dubois, P. J. (2014). Application of the overclaiming technique to scholastic assessment. *Educational and Psychological Measurement*, *74*(6), 975–990. https://doi.org/10.1177/0013164414536184

Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence*, *32*(3), 297–314. https://doi.org/10.1016/j.intell.2004.02.001

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

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, *60*(2), 307–317. https://doi.org/10.1037/0022-3514.60.2.307

Phillips, D. L., & Clancy, K. J. (1972). Some effects of "social desirability" in survey studies. *American Journal of Sociology*, *77*(5), 921–940. https://doi.org/10.1086/225231

Piotrowski, J. T., Litman, J. A., & Valkenburg, P. (2014). Measuring epistemic curiosity in young children: Brief report. *Infant and Child Development*, *23*(5), 542–553. https://doi.org/10.1002/icd.1847

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research*, *49*(6), 518–543. https://doi.org/10.1080/00273171.2014.933762

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Raubenheimer, A. S. (1925). An experimental study of some behavior traits of the potentially delinquent boy. *Psychological Monographs*, *34*, 1–107.

Reynolds, M. R., Keith, T. Z., Flanagan, D. P., & Alfonso, V. C. (2013). A cross-battery, reference variable, confirmatory factor analytic investigation of the CHC taxonomy. *Journal of School Psychology*, *51*(4), 535–555. https://doi.org/10.1016/j.jsp.2013.02.003

Reynolds, M. R., Keith, T. Z., Ridley, K. P., & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence*, *36*(3), 236–260. https://doi.org/10.1016/j.intell.2007.06.003

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037//1082-989X.7.2.147

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). 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

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, *2*(3), 114–121. https://doi.org/10.1111/j.1751-228X.2008.00041.x

Schneider, W., & Löffler, E. (2016). The development of metacognitive knowledge in children and adolescents. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory*. Oxford University Press.

Schreiber, A., & Marcus, B. (2020). The place of the "Dark Triad" in general models of personality: Some meta-analytic clarification. *Psychological Bulletin*, *146*(11), 1021–1041. https://doi.org/10.1037/bul0000299

Schroeders, U., & Gnambs, T. (2020). Degrees of freedom in multigroup confirmatory factor analyses: Are models of measurement invariance testing correctly specified? *European Journal of Psychological Assessment*, *36*(1), 105–113. https://doi.org/10.1027/1015-5759/a000500

Schroeders, U., Schipolowski, S., & Wilhelm, O. (2020). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 5. Bis 7. Klasse (BEFKI 5-7)*. Hogrefe Verlag.

Schroeders, U., Schipolowski, S., Zettler, I., Golle, J., & Wilhelm, O. (2016). Do the smart get smarter? Development of fluid and crystallized intelligence in 3rd grade. *Intelligence*, *59*, 84–95. https://doi.org/10.1016/j.intell.2016.08.003

Steger, D., Schroeders, U., & Wilhelm, O. (2021). Caught in the act: Predicting cheating in unproctored knowledge assessment. *Assessment*, *28*(3), 1004–1017. https://doi.org/10.1177/1073191120914970

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Vonkova, H., Papajoanu, O., & Stipek, J. (2018). Enhancing the cross-cultural comparability of self-reports using the overclaiming technique: An analysis of accuracy and exaggeration in 64 cultures. *Journal of Cross-Cultural Psychology*, *49*(8), 1247–1268. https://doi.org/10.1177/0022022118787042

Watts, L. L., Kuzmich, I., Leung, D. W., Gibson, C., & Barsa, A. (2021). Impression management and overclaiming on job applications: Related to future sales performance? *International Journal of Selection and Assessment*, *29*(1), 134–138. https://doi.org/10.1111/ijsa.12310

Weiss, S., Steger, D., Kaur, Y., Hildebrandt, A., Schroeders, U., & Wilhelm, O. (2020). On the trail of creativity: Dimensionality of divergent thinking and its relation with cognitive abilities, personality, and insight. *European Journal of Personality*, *35*, 22. https://doi.org/10.1002/per.2288

Zettler, I., Thielmann, I., Hilbig, B. E., & Moshagen, M. (2020). The nomological net of the HEXACO model of personality: A large-scale meta-analytic investigation. *Perspectives on Psychological Science*, *15*(3), 723–760. https://doi.org/10.1177/1745691619895036

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The vocabulary and overclaiming test (VOC-T). *Journal of Individual Differences*, *34*(1), 32–40. https://doi.org/10.1027/1614-0001/a000093

Zimprich, D., Nusser, L., & Pollatos, O. (2020). Are interoceptive accuracy scores from the heartbeat counting task problematic? A comment on Zamariola et al. (2018). *Biological Psychology*, *152*, 107868. https://doi.org/10.1016/j.biopsycho.2020.107868

IV Manuscript III

Stop Worrying About Multiple-Choice: Fact Knowledge Does Not Change with Response Format

The experimental idea and the study design were developed by Prof. Dr. Oliver Wilhelm and myself. I compiled and programmed the test battery for both studies reported in the manuscript with the assistance of Catherine Schittenhelm and Marlena Staab. The organization of the data collection was conducted by myself. The data analysis and discussion of results were conducted by myself under the supervision of Prof. Dr. Oliver Wilhelm. I have completed a literature search and drafted the original manuscript. All co-authors revised the manuscript under my supervision as first author. The manuscript was accepted for publication in the *Journal of Intelligence* on the 9th November 2022. The article was published online and open access on 14th November 2022 and falls under the following copyright: Open Access CC BY 4.0, https://creativecommons.org/licences/by/4.0

Goecke, B., Staab, M., Schittenhelm, C., & Wilhelm, O. (2022). Stop Worrying about Multiple-Choice: Fact Knowledge Does Not Change with Response Format. *Journal of Intelligence*, *10*, 1-22. https://doi.org/10.3390/jintelligence10040102

Journal of
Intelligence

MDPI

Article

Stop Worrying about Multiple-Choice: Fact Knowledge Does Not Change with Response Format

Benjamin Goecke *, Marlena Staab, Catherine Schittenhelm and Oliver Wilhelm

Institute for Psychology and Pedagogy, Ulm University, Albert-Einstein-Allee 47, 89081 Ulm, Germany
* Correspondence: benjamin.goecke@uni-ulm.de

Abstract: Declarative fact knowledge is a key component of crystallized intelligence. It is typically measured with multiple-choice (MC) items. Other response formats, such as open-ended formats are less frequently used, although these formats might be superior for measuring crystallized intelligence. Whereas MC formats presumably only require recognizing the correct response to a question, open-ended formats supposedly require cognitive processes such as searching for, retrieving, and actively deciding on a response from long-term memory. If the methods of inquiry alter the cognitive processes involved, mean-changes between methods for assessing declarative knowledge should come along with changes in the covariance structure. We tested these assumptions in two online studies administering declarative knowledge items in different response formats (MC, open-ended, and open-ended with cues). Item difficulty clearly increases in the open-ended methods although effects in logistic regression models vary slightly across items. Importantly, latent variable analyses suggest that the method of inquiry does not affect what is measured with different response formats. These findings clearly endorse the position that crystallized intelligence does not change as a function of the response format.

Keywords: crystallized intelligence; response formats; declarative knowledge

check for updates

Citation: Goecke, Benjamin, Marlena Staab, Catherine Schittenhelm, and Oliver Wilhelm. 2022. Stop Worrying about Multiple-Choice: Fact Knowledge Does Not Change with Response Format. *Journal of Intelligence* 10: 102. https://doi.org/10.3390/jintelligence10040102

Received: 30 July 2022
Accepted: 9 November 2022
Published: 14 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Fact knowledge questions are frequently used to assess students' learning progress and serve as a default procedure for assessing declarative knowledge as a pivotal component of crystallized intelligence (Gc; c.f. Wilhelm and Kyllonen 2021). Prototypically, such questions are administered in multiple-choice (MC) formats. However, alternative open response formats are regularly suggested to overcome possible disadvantages of MC response formats such as fostering recognition in contrast to actual retrieval of information from long-term memory (LTM) (e.g., Becker and Johnston 1999). In fact, open-ended response formats have been argued to measure distinct cognitive processes as compared to what is measured with MC formats (e.g., Becker and Johnston 1999; Hickson and Reed 2011; Krieg and Uyar 2001), but the evidence regarding this supposition is elusive (e.g., Scully 2017) and should be brought into perspective.

With the current studies, we contrast three competing response formats for assessing declarative knowledge as understood by extensions of contemporary intelligence structure models (Carroll 1993; McGrew 2005, 2009; Wilhelm and Kyllonen 2021). Study 1 aimed at contributing to a better understanding of how different response formats affect the empirical difficulties of conventional fact knowledge items and to what extent mean-changes in empirical difficulties between response formats of items otherwise equivalent occur. In addition to that, in Study 2 we explored the consequences of different response formats on the covariance structure of the tests from a multi-trait multi-method perspective of individual differences (Campbell and Fiske 1959). In the next sections, we embed declarative knowledge into the framework of crystallized intelligence, describe different modes of its measurement, discuss their advantages and disadvantages in application,

and further describe the possible underlying cognitive processes required by the different response formats.

1.1. Declarative Knowledge as an Indicator of Crystallized Intelligence

The study of knowledge complements efforts to understand and explain individual differences in human cognitive abilities (Cattell 1957), and especially crystallized intelligence (Gc). Gc is "typically described as a person's breadth and depth of acquired knowledge of the language, information and concepts of a specific culture, and /or the application of this knowledge" (McGrew 2009, p. 5). Although a broad Gc factor is proposed in contemporary intelligence structure models (Carroll 1993; McGrew 2005, 2009), it was argued that conventional measurements of Gc fail to include assessments of fact knowledge, although fact knowledge should be understood as a key component of Gc (Wilhelm and Kyllonen 2021). Knowledge regarding broad and general content domains, both curricular and extracurricular (Cattell 1971; Schipolowski et al. 2014; Wilhelm and Kyllonen 2021), should also be deemed the result of ongoing acculturation processes (Ackerman 2000) and of intellectual investment traits (e.g., von Stumm and Ackerman 2013).

Tests of fact knowledge have been shown to load highly on a general factor of Gc, which was marked by verbal and language-related abilities (e.g., Horn 1965). It was argued that declarative knowledge tests under the consideration of various content domains can serve as a good marker for Gc, in addition to existing measurement efforts focusing on language-related abilities (Ackerman 2000; Amthauer et al. 2001). In fact, latent factors for verbal abilities and fact knowledge were found to be correlated near unity (Schipolowski et al. 2014), which supports this view. Indeed, measurement instruments used for indicating Gc increasingly employ broadly sampled tests of fact knowledge André (Beauducel and Kersting 2002; Schipolowski et al. 2014; Schroeders et al. 2020), depicting a broad factor of Gc covering knowledge that individuals accumulate during their lifetime through acculturation and learning (Cattell 1971, 1987).

Recent studies investigating the psychometric structure of Gc suggest a higher-order model, with a general factor of declarative knowledge capturing the strong positive manifold of broadly sampled content domains such as natural sciences, life sciences, humanities, and social sciences modeled as first-order factors (Steger et al. 2019). This is in line with the view that it is pivotal to conduct a broad assessment of declarative knowledge rather than administering a single in-depth and domain-specific test (Ackerman 1996, 2000). Taken together, Gc should therefore best be measured with adequate broadness with respect to its knowledge domains (Ackerman 2000; Steger et al. 2019).

1.2. Assessment Methods of Declarative Knowledge

Measuring declarative knowledge seems simple: administer fact knowledge questions which refer to knowledge that persons can actually acquire during their lives (Wilhelm and Schroeders 2019). The basic assumption of administering such questions is that they tap into what persons can in principle know. Apart from bearing in mind individual sample characteristics such as age (Li et al. 2004; Watrin et al. 2022), education levels (Ackerman 1996), or even characteristics of specific item samples, such as item samples suited especially for children in primary school (Schroeders et al. 2016), research on overarching psychological dispositions should preferably transcend the methods of inquiry we apply. That is, Gc should go beyond the specificities of a single test and one important way to abstract from specificities is to vary presumably irrelevant attributes of measures. It is an open question whether or not the response format of fact knowledge questions is irrelevant in the sense that they might change what a fact knowledge test measures.

Generally, two response modes for measuring Gc (and other abilities) can be distinguished: selected-response formats such as multiple-choice (MC) formats and so-called constructed-response (CR) formats (e.g., Traub 1993). Both types differ with regard to the format in which participants respond, but both depend on an item-stem that articulates a stimulus (usually a question) awaiting a response. In the case of declarative knowledge,

the item-stem is usually a question with only one veridical solution. Whereas the MC format provides several response alternatives per question, of which the correct one has to be chosen (Downing and Haladyna 2006), the CR format does not provide response alternatives and usually requires test takers to retrieve the correct response from memory (e.g., Martinez 1999; Sam et al. 2018, 2019; Traub 1993). Other tests based on CR formats can require subjects to build a response by assembling given pieces of information (e.g., certain reasoning tasks Becker et al. 2015).

1.3. MC Format Items vs. CR Format Items

MC-format items are used for measuring a broad variety of abilities in many and diverse applied test settings (e.g., SAT since 1937, Hancock 1994; university exams, Lindner et al. 2021; PISA, Rodriguez 2003; driving license tests) since their first large scale applications through Robert Yerkes in the Army Alpha Intelligence Tests during World War I (Chan and Kennedy 2002).

Their popularity is indicative of their inherent advantages (Chan and Kennedy 2002; Walstad 2001): most saliently cost-efficient development (Abu-Zaid and Khan 2013) and administration (both paper-pencil and computerized; Schipolowski et al. 2014; Watrin et al. 2022, respectively); easily automated scoring that is not subject to reliability issues inherent in many other response formats (e.g., Kennedy and Walstad 1997); and finally, administration with simple and swift-to-deliver instructions due to the intuitive response format is easy and does not require costly training of proctors. In turn, these advantages of MC-format items can be understood as the disadvantages of CR formats (e.g., Chan and Kennedy 2002).

MC tests in general have been shown to yield good reliabilities and validity (Downing and Haladyna 2006) and are deemed "the workhorse of the testing enterprise throughout much of the world" (Downing and Haladyna 2006, p. 293). We suggest that this statement also pertains to MC tests assessing declarative knowledge (e.g., Schroeders et al. 2016; Steger et al. 2019). However, there are also reservations regarding MC-format tests and some studies set out to measure fact knowledge solely relying on CR item formats (Lynn et al. 2001; Lynn and Irwing 2002).

One concern regarding MC tests is that they might facilitate learning wrong information due to uncorrected errors, for example in college exams or other application contexts (Fazio et al. 2010). MC tests have also been argued to encourage guessing (Becker and Johnston 1999). The ability to infer the correct response from hints such as the provided response alternatives, or at least to eliminate individual distractors in the presence of incomplete expertise can also be considered an aspect of "test wiseness" or "test-taking ability" (Millman et al. 1965; Sabers 1975). If test takers are test-wise and possess relevant knowledge and if a test contains susceptible items, then the combination of these factors will result in higher scores; responses on MC-format items can be inferred without actually providing a question and these inferences seem to be driven by working memory, and prior knowledge (Hartung et al. 2017). Supposedly, strategies to take MC tests are widely established among test takers (Chittooran and Miles 2001; Kesselman-Turkel and Peterson 2004).

Questions about construct-equivalence across different response formats such as MC and CR are the subject of discussion (Rodriguez 2003). In contrast to the MC format, CR formats are often used in test settings, because they supposedly allow measuring more complex cognitive processes (such as applying actual knowledge or evaluating a concept, c.f. Bloom's taxonomy, Krathwohl 2002) more readily than the MC format does (Martinez 1999; Rodriguez 2003). Supposedly, this view is supported by studies showing that test scores based on MC-format items only explained 50% of the variance in test scores based on essay-like CR-format items (Hickson and Reed 2011). Further studies either found no support for a single-factor solution across both response formats (Traub and Fisher 1977) or found no relationship at all between both response formats (Becker and Johnston 1999).

Opposed to these findings, several studies found MC-format tests and CR-format tests to be equally well indicative of one and the same underlying construct. For example, verbal and reading skills have been found to be best described by a single latent factor, indicating

the high communality across test formats (Vernon 1962; Ward 1982). In addition, there is evidence for construct-equivalence between MC and CR formats in mathematical reasoning items (Traub and Fisher 1977). Further studies of numerical problem solving also revealed very high correlations (r = .90) between response formats (Harke et al. 1972; Horn 1966). Regarding knowledge questions, it was also shown that CR formats add little information beyond MC-format items (Lukhele et al. 1994; Thissen et al. 1994). This is supported by meta-analytical evidence for construct equivalence between response formats of MC and CR items, which also shows that if the same item-stem was used for investigating correlations between response formats, unity was almost approached (Rodriguez 2003). Lastly, evidence from two large-scale educational assessments (TIMSS and PIRLS; Mullis et al. 2007, 2008) shows that differences between the response formats with regard to their reliability and validity can be neglected (Schult and Sparfeldt 2018).

1.4. Cognitive Processes Underlying Different Response Formats

Responses to declarative knowledge questions are arguably not simply driven by crystallized information structures, but also through a variety of retrieval processes. These retrieval processes might vary across response formats. The correct answer to a knowledge question has to be percolated from existing information in long-term memory (e.g., Tulving and Watkins 1973; Unsworth 2019), that is, it has to be found amongst a manifold of other irrelevant information. For a knowledge question in MC format, one could argue that a response builds upon *recognition* of the correct response option. Actively retrieving information from long-term memory is required in CR response formats and this cognitive activity is best-labeled *recall* (e.g., Haist et al. 1992). Many authors argue, that CR response format items are needed to measure more complex cognitive processes such as a new combination of existing information in memory (e.g., Martinez 1999), and that MC-format items are not suited to assess more than pure recognition (Martinez 1999; Veloski et al. 1999).

The discontinuity hypothesis states that recognition and recall are two fundamentally different memory processes (Tulving and Watkins 1973). For example, recognition might be understood as one phase of the identification of a response, whereas recall requires two phases, namely an exhaustive search for a response and a proactive decision for it (Anderson and Bower 1972; Gillund and Shiffrin 1984). According to the discontinuity hypothesis, it is questionable whether two latent variables for declarative knowledge measured with either MC or CR format are strongly related. Presumably, both factors capture performance in different memory processes and do not reflect the same underlying ability.

Whether MC-format items and CR-format items measure the same underlying ability equally well, might depend on the ability that is to be measured, and on the actual format of the administered items (Hancock 1994). Accordingly, Tulving and Watkins (1973, p. 739) stated: "A critical problem of long standing in the psychological study of memory is concerned with the relation between recall and recognition. In what sense are they the same, and in what sense are they different?". This question can be readily transferred to the study of competing response formats of declarative knowledge, such as MC formats and CR formats.

In an MC-format test with one correct response and three distractors, the guessing probability is 25%. If the distractors are additionally implausible, simply falsifying all implausible response alternatives might facilitate answering an item. This rejuvenates the question of whether the test takers of MC-format tests actually know the answer to a given item stem or if they simply recognize it after considering the incorrect distractors. If MC-format tests reward partial knowledge (i.e., recognition) just like real in-depth knowledge (i.e., recall), then test takers might learn to discard implausible or plainly wrong distractors instead of effortfully verifying the correct response from memory (Martinez 1999; Scouller 1998). Obviously, triggering such elimination strategies is not the measurement intention when MC format is used, and the degree to which this account for MC-format performance applies is a controversial topic (Coleman et al. 2010; Daneman and Hannon 2001;

Fowler and Kroll 1978; Katz et al. 1990, 1991; Martinez 1999; Rost and Sparfeldt 2007; Sparfeldt et al. 2012).

Whereas MC response formats offer cues as to what the correct response to a knowledge question is (it has to be one of the four provided response alternatives), CR response formats usually do not present any cues regarding the correct response other than the question itself. However, the degree to which cues in open-ended response formats are presented can be manipulated experimentally. For example, instead of not restricting responses to a CR question at all, one could provide the first letter of the correct answer. This would limit the number of possible responses to the question and facilitate the executed retrieval processes (i.e., an exhaustive search of possible responses). With respect to pure mean-structure effects, for which one stable result is that MC-format items usually show higher solution probabilities (e.g., Chan and Kennedy 2002; Hohensinn and Kubinger 2011; Sam et al. 2018, 2019), the mean performance of such a cued open-ended response format should be located between conventional MC-format items and open-ended format items.

Compared to the MC format as the prototypical representative of selected-response formats, the CR format comes with more degrees of freedom regarding test construction. Whereas MC formats might hedge one pole of a fictional response format continuum, the CR format would cover the rest of the continuum in different manifestations, spanned from cued open-ended response formats that only require responding with one word to open-ended response formats that are akin to essays. Depending on the location of a response format on the continuum, the requirement of different cognitive processes for responding to an item-stem might change.

From a psychometric point of view, measuring declarative knowledge with distinct response formats (e.g., MC formats vs. open-ended formats), can plausibly deliver three alternative outcome scenarios: First, administering different response formats to measure the same underlying ability affects neither the mean structure of the items and thus the test, nor its covariance structure and thus what such tests measures in the first place. Second, the mean structure of items hinges upon the response format, but the covariance structure (i.e., the rank-order of individual test takers) is unaffected, which can indicate that the same underlying ability is tapped by the test, independent of the used response format. Third, in addition to the mean structure, the covariance structure is affected, which would indicate that different response formats measure different underlying cognitive abilities (e.g., Hickson and Reed 2011; Scully 2017).

1.5. The Present Studies

Studying declarative knowledge implicitly follows a seemingly longstanding tradition of using MC item formats (Ackerman 1996; Steger et al. 2019). With the present studies, we aimed at addressing the following questions: How do different response formats affect item difficulty? Do changes in the response format cause changes in the rank-order of individuals? Is the same ability measured across different response formats? We addressed these questions by investigating declarative knowledge items in three competing response formats.

To this end, we administered a broad test battery of declarative knowledge items across the following response formats: MC-format items with four response alternatives, an open-ended CR format, and an open-ended CR format with cues (e.g., the first letter of the correct response).

In study 1, we tested how empirical item difficulties of a Gc test are affected by response formats. We hypothesized that other things being equal MC items should be easiest because they supposedly rely on recognition of the correct response only rather than actually retrieving it from long-term memory. The cued open-ended response format should be more difficult, because it may require a more sophisticated retrieval process when mentally searching for the correct response to a knowledge question. Lastly, the open-ended item format should be more difficult to give a correct response to, because the room for possible responses is only restricted through the content of the knowledge question. The

proposed mean effects should be apparent in both comparisons of the solution probability of single items, and aggregate scores of the response formats.

Importantly, it is unclear whether the response formats tap the same underlying cognitive ability. Arguably, different response formats require distinct cognitive processes (e.g., *recognition* in MC formats vs. *recall* in open-ended formats), and if this is the case, this should affect the covariance structure of tests, and thus the rank-order of individuals across response formats. In study 2, we tested for changes in covariance structure across response formats. To this end, we evaluated a series of competing measurement models. The research objectives and hypotheses were not preregistered.

2. Methods and Materials: Study 1

2.1. Participants and Procedure

Study 1 was conducted online (using *SociSurvey*). The study was advertised via mailing lists of the local university, via social media, and via a German online panel (*Respondi*). We conducted the study commensurate with the ethical guidelines provided by the German Society for Online Research (*DGOF*) and the EU General Data Protection Regulation (*GDPR*). The anonymity of the participants was guaranteed. Ethical approval was not required as per local legislation. In the online panel, participants were reimbursed according to the reimbursement rules of *Respondi*.

In total, $N = 198$ participants completed study 1. In order to ensure data quality, participants were excluded from data analysis, if they (a) indicated to have used unpermitted aids (e.g., using Google search during the study) ($n = 36$), (b) indicated to have participated without care ($n = 16$) or failed attention checks ($n = 1$). In addition, participants were removed if they were identified as outliers regarding defocusing events (i.e., changing tabs in their browsers >3 *SD*; $n = 4$). The final sample consisted of $N = 142$ participants. The mean age of the sample was $M = 29.8$ years ($SD = 6.97$ years, range = 18–37), and 45.8% were female. Approximately 49% of the sample indicated holding at least a high school degree.

2.2. Measures

Declarative Knowledge. In study 1, we assessed declarative knowledge using 72 items covering four broad knowledge domains and twelve subdomains, namely natural sciences (physics, biology, chemistry), social sciences (politics, law, economy), humanities (art, literature, music), and life sciences (medicine, nutrition, health) (c.f., Steger et al. 2019). All items were sampled from a large item pool (Steger et al. 2019). The items were sampled to cover the above-mentioned knowledge domains. Please note, that the item pool of Steger et al. (2019) was developed by reviewing existing knowledge test batteries (see Steger et al. 2019 for an overview). Additionally, the authors of the item pool ensured that items from various vocational profiles and education contexts were included in the original item pool (see Steger et al. 2019). The items we sampled from this pool for our studies were administered in three response formats:

1. MC-format items included a knowledge question (e.g., "What is the capital of Sweden?") and four response alternatives with exactly one veridical response;
2. For open-ended format items, participants were only presented with the knowledge question and a text box for typing in the response;
3. The same was true for the cued open-ended format items, although in this particular response format, participants were additionally provided with a cue (the first letter of the correct response, or a restriction as to the range of the correct number).

Items were presented one at a time to the participants. All participants were instructed to guess the correct response in the MC format if they did not know the answer. Further, if they did not know an answer in the open-ended or cued format, they were instructed to indicate this by filling in "I do not know" or a question mark. There were no time restrictions and it was clearly stated in the instructions that participants should not use any aids for responding to the items.

All items were administered in all three response formats. All participants responded to all 72 items, but participants were randomly assigned to one of three item-sets, each consisting of a distinct mixture of items from different response formats so that each item in each response format was answered by approximately one-third of the sample (c.f., Table 1). No item was shown more than once to a participant.

Table 1. Design of the studies.

Study	Group (*N*)	*n* (Items) + Itemset			Total Items
		MC	Cued Open-Ended	Open-Ended	
1	1 (*N* = 46)	24 A	24 C	24 B	72
	2 (*N* = 50)	24 B	24 A	24 C	72
	3 (*N* = 46)	24 C	24 B	24 A	72
2	1 (*N* = 300)	24	24	24	72

Note. The reported sample sizes are after data exclusion. Each item set of Study 1 (a, b, c) contained different items and the item sets were allocated to different response formats across the groups of Study 1.

Some items were piloted in different response formats (see Supplementary Materials, p. 1). Descriptive statistics for all items in all response formats across both studies (and our selection of the item pool of Steger et al. (2019)) are provided in the Supplementary Materials (SM Table S1).

Para Data—Defocusing. During testing, JavaScript tracked whether or not participants changed tabs or windows (similar to Diedenhofen and Musch 2017; Steger et al. 2020). The change in a tab is thought to capture the occurrence of so-called defocusing events and was found to be predictive of cheating in unproctored test settings (Steger et al. 2020). However, defocusing (i.e., changing a tab during an online test) should not be equated with cheating behavior per se. Although changing a tab during an online test session might be made to cheat, not all tab changes must indicate cheating. Further, cheating is also possible without changing tabs, for example through using a second digital device.. Therefore, we added this count variable as a covariate to all of our analyses in order to demonstrate that our results and main conclusions are barely affected by defocusing events such as tab changes. Nevertheless, we excluded participants with large amounts of defocusing events, because we take this as an indicator of inattentive responding (see Section 2.1 *Participants and Procedure* above).

2.3. Data Preparation

Prior to the main statistical analysis and after initial data cleaning, the data of the items administered in the open-ended and cued response formats were scored by human raters. All items were scored dichotomously. Items were scored as correct if participants indicated either the response which was deemed correct in the original MC item format as provided by the original item pool (Steger et al. 2019), or if participants indicated an alternative but still correct response (which was only possible in the open-ended response formats). Obvious spelling mistakes such as forgetting a letter were not counted as incorrect, whereas wrong names, wrong formulas, and answers that indicated not knowing the correct response were dismissed.

Two raters scored all responses for all persons independently. Beforehand, the raters were provided with acceptable and expected responses for each item. Cohen's κ (Kappa; Cohen 1960) was used to determine the inter-rater reliability. After initial scoring, the average κ's across the open-ended and the cued items were $M_{\kappa(\text{open})} = .97$ ($SD_{\kappa(\text{open})} = .07$) and $M_{\kappa(\text{cued})} = .97$ ($SD_{\kappa(\text{cued})} = .07$), respectively, indicating minor deviations of the two

scorers. These minor deviations could be clarified, so that the final scoring of the open-ended data yielded in full agreement of the scorers ($M_{\kappa(open)} = 1$ and $M_{\kappa(cued)} = 1$).

2.4. Statistical Analysis

All analyses were conducted with R (R Core Team 2022). To make all analyses reproducible, we provide all materials necessary, including data and analysis scripts, in an online repository: https://osf.io/pse3w/.

In order to compare the difficulties of the items between response format conditions, we built means for all item-sets in all response format conditions and computed Cohen's *d* as a measure of effect size (Cohen 1969). Additionally, in study 1 we used logistic mixed regression models. The items of each response format provide a dichotomous independent variable indicating correctness. Thus, we analyzed our data by means of logistic regression models, in which the correctness of any given item was predicted by response format (i.e., MC vs. cued open-ended vs. open-ended). In addition to the fixed effect of this predictor, we specified a full model that included the main effect of defocusing to account for participants that may have been cheating, and the interaction term of response format and defocusing to account for potential differences in tab-changing dependent on the response format, a random effect of the participants (i.e., random intercept), and a random effect of the item (i.e., random intercept), because certain facts might be better known than others. As such, we accounted for the multi-level data structure. This full model was compared with more parsimonious models (i.e., dropping single effects), to evaluate each effect. Regression weights of the main effect s can be interpreted relative to the respective reference response format.

3. Results: Study 1

In Figure 1, we report the mean difficulties of response formats for each itemset. In line with expectations, these statistics indicate that the MC format is easiest in each of the administered item samples. The means within one response format across item-sets are highly similar. The effect sizes for the mean differences between the response formats range between $d = .72–1.10$ for MC compared to the cued open-ended response format, $d = 1.31–1.49$ for MC compared to the open-ended response format, and $d = .33–.57$, for the cued open-ended compared to the open-ended response format, respectively. Almost all mean differences were statistically substantial (all $p < .05$; the exact test statistics can be found in the Supplementary Materials SM Table S3), with the exception of the mean difference between the cued open-ended and the open-ended response format in itemset B.

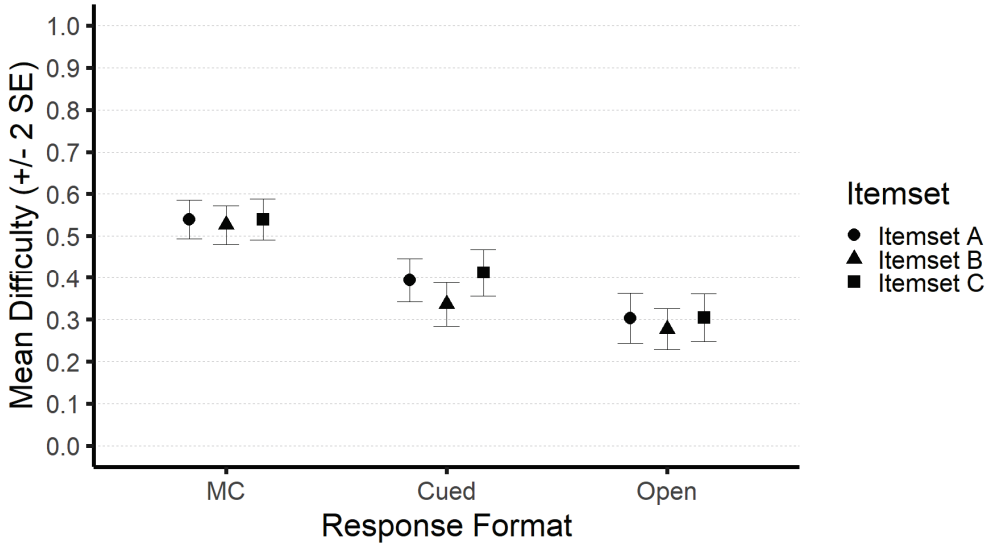


Figure 1. Mean difficulties of response formats for each item set.

Next, we compared several logistic regression models that used the MC response format as a reference method, so that the regression weights can be interpreted relative to items of the respective reference method. We tested the full model, as described above, against two more parsimonious models that did not include (a) the random intercept for single items, and (b) the random intercept for participants. In both cases, a chi-square difference test revealed that dropping either one of the effects deteriorated model fit substantially (a: $\Delta\chi^2(1) = 1580.9$, $p < .001$, and b: $\Delta\chi^2(1) = 799.1$, $p < .001$, respectively). We thus kept both effects in the model. Additionally, we tested the full model against a model without the interaction term between response format and defocusing, an effect that implies differential effects of defocusing depending on the response format. Again, a chi-square difference test revealed that the full model fitted the data somewhat better ($\Delta\chi^2(2) = 9.56$, $p < .01$). Therefore, we also kept this interaction effect, in addition to the main effects of response format and defocusing.

Parameter estimates of the final model are presented in Table 2. The main effects are in line with expectations. Solving MC items is easier than solving cued open-ended or open-ended items. The odds in favor of solving a cued open-ended item relative to an open-ended item were less pronounced. These effects persisted if defocusing events are included in the model. The main effect of defocusing is substantial (indicating an increased probability of solving an item when defocusing events occur; c.f., Steger et al. 2020). Importantly, with and without defocusing MC format is easiest, followed by cued open-ended which in turn is easier than open-ended format. Although the fixed effect alone explained little variance in the criterion (marginal R2 = .083; Nakagawa et al. 2017), the effects of the response format were still substantial. Across the test battery of study 1, we found evidence that MC items are more frequently solved as compared with their (cued) open-ended counterparts.

Table 2. Logistic regression model with standardized regression weights and odds ratios (OR).

Model		With Defocusing		
MC Reference	**Fixed Effects**	**Estimate (*SE*)**	***OR***	***p***
	(Intercept)	.22 (.15)	1.25	.15
	Cued	−.98 (.06)	.38	<.001
	Open	−1.55 (.07)	.21	<.001
	Defocusing	.65 (.18)	1.91	<.001
	Defocusing*Cued	.24 (.21)	1.28	.24
	Defocusing*Open	.61 (.21)	1.85	<.01
	Random effect	Variance (*SD*)		
	(Intercept of Person)	.76 (.87)		
	(Intercept of Item)	1.20 (1.10)		

An odds ratio of 1 indicates no difference between formats, whereas odds ratios > 1 indicate that the likelihood of a participant correctly responding to an item increases, relative to the reference group. Analogously, odds ratios < 1 indicate that the likelihood of a participant correctly responding to an item decreases, relative to the reference group. Please note that three items were excluded from these analyses because they were declared as outliers regarding their odds as compared to at least one of the other response formats (they exhibited ORs >3 SD from the mean distribution of ORs). Odds ratios for single items are provided in Supplementary Materials (SM Table S2).

4. Discussion: Study 1

We administered 72 fact knowledge items from four broad knowledge domains (Steger et al. 2019) in three response formats each. This approach allowed for a thorough test of item difficulty as a function of response format. In line with previous research, we hypothesized that different cognitive requirements might be required by more open response formats (i.e., simple *recognition* vs. more complex *recall* from long-term memory), and should be associated with lower probabilities of solving an item correctly. In other words, if surface

characteristics in a diverse set of items matter, they should account for individual differences in subsets of items once the general ability is controlled for. In line with these predictions, we observed the hypothesized difficulty order of item formats. Across most items, and across all used item-sets, we found that MC-format items exhibit the highest solution probabilities, followed by the cued open-ended and concluded by the open-ended format. This rank-order might be explained by the different cognitive processes that underly solving an item of a respective response format. Whereas MC-format items naturally offer more readily accessible information through presenting test takers with alternative responses, this is not the case in the open-ended response formats—as a consequence, it is harder to solve an item in the open-ended response format and the solution probability of an item is directly associated with the a priori provided information.

It should be noted that the magnitude of all effect size estimates, on the item level, but also on the score level (that is, aggregated scores across items per participant), were relatively large, and thus the mean differences between the response formats can be deemed fairly general and stable, and cannot be easily attributed to statistical artifacts such as differences in guessing probabilities. Arithmetically, the guessing probability of the administered MC-format items was .25 per item. In contrast, the guessing probabilities of the open-ended and cued open-ended items should be approximately zero due to the inherent nature of these item formats. The arithmetic guessing of the probability of MC-format items rests on the strong assumption that all distractors (i.e., erroneous response choices) are equally attractive—however, this assumption hardly ever holds, even if item development follows the strongest quality benchmarks. As knowledge accumulation is an idiosyncratic process, the attractiveness of a distractor might also depend on person-bound characteristics, and as such, the guessing probability for MC-format items can only hardly be accounted for. Although 3-parameter IRT-models might be fit to control for arithmetic guessing probabilities by estimating a specific parameter for it, these models require large sample sizes than our Study 1 provides and are objectionable for a number of reasons, including the above-mentioned (see, e.g., Chapelle 1999).

Nevertheless, we argue that further studying guessing effects to exclude this as an explanation for the here presented results is warranted. Empirically, guessing effects could be studied across different instantiations of knowledge items that are mixed within-subject administration (i.e., between-subjects). These instantiations could span from items in an open-ended response format, over a cued open-ended format, towards individually increasing the number of response alternatives for a given question (i.e., one attractor and one distractor; one attractor and two distractors, etc.). These data could then be used to obtain an approximate estimate of what effect any given arithmetic guessing probability has on a specific item pool.

5. Methods and Materials: Study 2

5.1. Participants and Procedure

Just like Study 1, Study 2 was conducted online (using *SociSurvey*) using the same online panel (*Respondi*). We determined the sample size for Study 2 a priori (*using simsem*; Pornprasertmanit et al. 2021), based on considerations of several parameter estimates for our target latent factor models (i.e., minimally expected effect sizes and typical loadings of indicators). Across various competing models, we found that a sample size of $N = 307$ yields sufficient power (>.80, $\alpha = .001$) for all targeted parameter estimates (including factor loadings and factor correlations). In total, $N = 376$ participants were included in study 2. Participants were excluded from data analysis if they (a) indicated to have used unpermitted aids ($n = 39$), (b) indicated to have participated without care ($n = 19$), or (c) failed attention checks ($n = 1$). Again, participants were removed if they changed tabs extensively during testing (>3 *SD*; $n = 7$). In addition, $n = 2$ persons encountered technical difficulties so no responses were saved in their data files. Lastly, after scoring and computing test scores, $n = 8$ participants were identified as multivariate outliers and thus removed from further analysis. The final sample size was $N = 300$. 70.3% of the sample

indicated to be female; the mean age was M = 29.53 years (SD = 7.04, range = 18–60). The majority of the sample (72.3%) indicated having at least a high-school degree.

5.2. Measures

Declarative Knowledge. In Study 2, we used the same 72 items for assessing declarative knowledge as in Study 1 but administered three item sets fixed to one response format each (i.e., 24 items per response format; c.f., Table 1). There was no item overlap between the response formats. Presentation for all 72 items, and hence the response formats, was randomized for each participant in order to prevent possible confounding through fixed item sequences. The broad knowledge domains were balanced in each item set., i.e., they were represented equally across the number of items and the response formats. For each response format, items were allocated considering the recommendation that declarative knowledge scales should cover a wide range of difficulty (Schneider and McGrew 2018). We thus chose items according to both their empirical difficulties as indicated through means and standard deviations of an MC response format item pool (N > 1000, Steger et al. 2019), and according to their empirical difficulties within the respective response formats of Study 1.

Para Data—Defocusing. For Study 2, we used the same procedure to detect defocusing events as described in Study 1 (see Section 2.1 *Participants and Procedure* of Study 1 above).

5.3. Data Preparation

We applied the same scoring procedure for the items of Study 2 as described for Study 1. The data of both open-ended response formats in Study 2 were also scored by several human raters. In contrast to Study 1, one person rated all open-ended items, whereas another person rated all cued open-ended items. In addition to that, a third person rated all items of both response formats. The inter-rater reliability after initial scoring was very high ($M_{\kappa(open)}$ = .99, $SD_{\kappa(open)}$ = .01; $M_{\kappa(cued)}$ = .99, $SD_{\kappa(cued)}$ = .02). After clarifying deviations full agreement for all items was reached.

5.4. Statistical Analysis

Again, all analyses were conducted with R (R Core Team 2022) and we provide all materials necessary to reproduce the analyses in an online repository: https://osf.io/pse3w/.

In Study 2, we aimed at comparing competing measurement models to address the dimensionality of Gc across response formats. We built parcels to reduce model complexity, after establishing the unidimensionality of the single response format scales (Little et al. 2002). We thus decided to exclude items with problematic item characteristics from the item sample. We first examined item difficulties and excluded MC items below the guessing probability for the MC format (i.e., .25). In addition, we examined the corrected item-test-correlations for all response formats and excluded items with r_{it} < .18. After that, we computed unidimensional measurement models per subscale and excluded items with non-significant factor loadings (*lambda*), or items which considerably deteriorated model fit (*misfit*). We then built four parcels per response format maintaining domain specificity according to the broad knowledge domains (Cole et al. 2016; Steger et al. 2019). In total, n = 8 items were excluded from subsequent analyses (see SM Table S5 for an overview of the measurement models on the item level). In addition, please note that we provide a full correlation matrix, including descriptive statistics, for all the indicators used in our measurement models in the Supplementary Materials (SM Table S4).

Confirmatory Factor Analysis (CFA) was carried out by the R package *lavaan* (Rosseel 2012). Where possible, we used *full information maximum likelihood* estimation under the assumption of missing completely at random to combine missing data and parameter estimation in a single step (Schafer and Graham 2002; Enders 2010). Models based on dichotomous indicators are based on the *Weighted Least Squares Means and Variance* adjusted (WLSMV) estimator (Beauducel and Herzberg 2006); in this case, pairwise observations were used. Models based on continuous indicators are based on a maximum likelihood

estimator with robust standard errors (MLR). The following fit statistics were considered to indicate good model fit: CFI (*Comparative Fit Index*) $\geq$ .95, RMSEA (*Root Mean Square Error of Approximation*) $\leq$ .06, and SRMR (*Standardized Root Mean Square Residual*) $\leq$ .08 (Hu and Bentler 1999). For acceptable model fit these boundaries were used: CFI $\geq$ .90, RMSEA $\leq$ .08, and SRMR $\leq$ .10 (Bentler 1990; Browne and Cudeck 1992). We used McDonald's ω as an indicator of factor saturation (McDonald 1999; Raykov and Marcoulides 2011). The factor saturation of a factor indicates how much variance is accounted for by a latent variable in all underlying indicators (Brunner et al. 2012). We specified all measurement models with and without the defocusing covariate (please see the Supplementary Materials SM Figures S1–S4).

6. Results: Study 2

Modeling Declarative Knowledge and Accounting for Response Formats

We compared competing measurement models to address the dimensionality of Gc across response formats (see Table 3). The first model specified correlated factors with one latent factor per response format (Figure 2; model A in Table 3). The model fit was acceptable. The correlation between Open and Cued was estimated slightly above unity indicating extreme collinearity between the response formats. In this model, individual differences in Gc are modeled as being due to response formants exclusively. The model negates the existence of knowledge domains and allows for different rank-orders of subjects across response formats. This model shows extremely high multi-collinearity between the different response formats, which shows that the rank order of subjects does not change across response formats. This model was extended with a manifest defocusing count variable as a predictor of the correlated group factors ($\chi^2(60) = 164.02$, CFI = .914, RMSEA = .076, SRMR = .045; c.f., SM Figure S1) and found that defocusing was a positive predictor of all three factors (β_{MC} = .28 (*SE* = .06), β_{Open} = .36 (*SE* = .04), β_{Cued} = .31 (*SE* = .07), respectively).

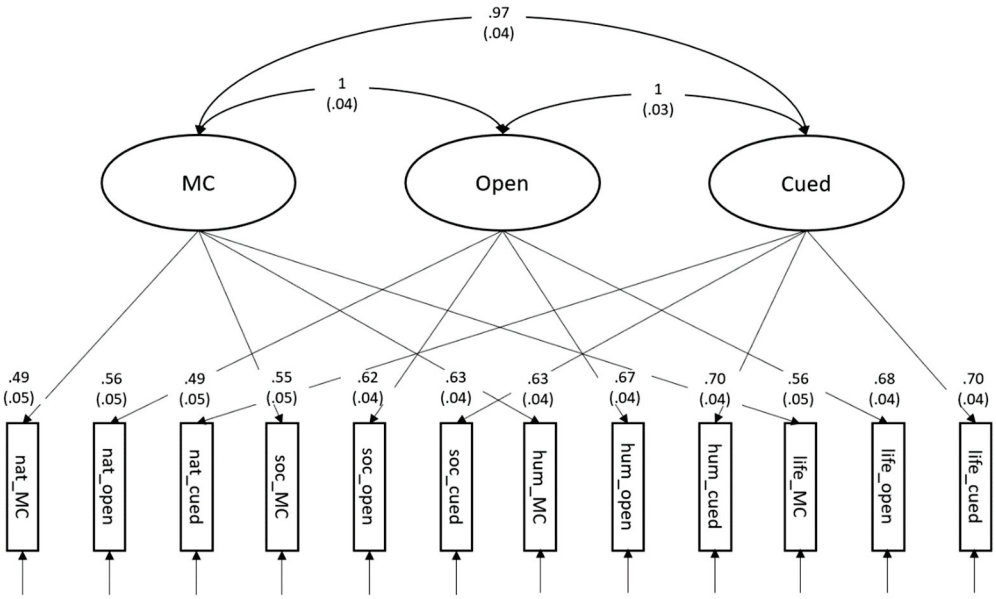


Figure 2. Correlated factors model of the response formats. All parameters are standardized. n = 300; $\chi^2(51) = 148.70$, CFI = .917, RMSEA = .080, SRMR = .045. Standard errors are depicted in parentheses. Please see SM Figure S1 for the model with defocusing.

Next, we specified a model in which all indicators were subsumed below a single general factor (model B in Table 3), which corresponds with the idea that response formats play no role in the covariance structure of a declarative knowledge test. A χ^2-difference test, indicated that both models were not different ($\Delta\chi^2(3, N = 300) = 2.43$, p = .49). The general

factor captured substantial variance ($p < .001$). The fit of the model was only acceptable. We extended this model with a manifest defocusing count variable as a predictor of the general factor as well ($\chi^2(65) = 168.15$, CFI = .914, RMSEA = .073, SRMR = .046; c.f., SM Figure S2) and found that defocusing was a positive predictor of the general factor ($\beta = .32$ ($SE = .06$)).

Table 3. Competing measurement models.

	Measurement Model	χ^2	df	CFI	RMSEA	[90% CI]	SRMR
A	3 Correlated Factors Response Formats	148.7	51	.917	.080	[.065; .095]	.045
B	g-factor	150.90	54	.918	.077	[.063; .092]	.045
C	4 Correlated Factors Knowledge Domains	57.87	48	.992	.026	[.000; .048]	.026
D	Higher-Order Knowledge Domains	62.98	50	.989	.029	[.000; .050]	.029

Note. All models are based on $N = 300$ participants and on $n = 12$ indicators.

Next, we established a correlated factors model with four factors for the four broad knowledge domains (model C in Table 3). The four factors captured individual differences across response formats and within knowledge domains. The model negates the existence of individual differences due to response formats and also does not allow for an overarching Gc factor. Instead knowledge domains are specified as distinct but correlated entities that do not adhere to a hierarchy of cognitive abilities. Therefore, this model depicts individual differences in knowledge domains, which are independent of the administered response formats. This model fitted the data significantly better than a single general factor ($\Delta\chi^2(6, N = 300) = 82.23$, $p < .001$). All factors captured substantial shares of variance ($\omega_{Nat} = .55$, $\omega_{Soc} = .74$, $\omega_{Hum} = .78$., and $\omega_{Life} = .74$, respectively). Overall, the correlations between the group factors were large (all $r > .7$; $p < .001$), but did not reach unity. The model fit was good.

We added a manifest defocusing count variable to model C ($\chi^2(56) = 67.04$, CFI = .991, RMSEA = .026, SRMR = .028) as a predictor of all group factors (natural sciences, social sciences, humanities, life sciences) in order to control for defocusing events (c.f., SM Figure S3). The standardized regression weights were $\beta_{Nat} = .42$ ($SE = .07$), $\beta_{Soc} = .30$ ($SE = .07$), $\beta_{Hum} = .27$ ($SE = .07$), $\beta_{Life} = .24$ ($SE = .06$), respectively. Defocusing accounted for 17.4%, 9.2%, 7.5%, and 5.5% of the variance of the factors, respectively. The remaining variances of the residuals of the latent factors (i.e., individual differences after controlling for defocusing) were still significant (all $p < .001$), and correlations between factors remained unaffected. This model is included in Supplementary Materials (SM Figure S3).

Next, we established a higher-order factor model with a general factor capturing the covariance of the four latent trait factors (natural sciences, social sciences, humanities, and life sciences; see model D in Table 3 and SM Figure S4). A higher-order factor explicitly represents the overarching ability (Gc), as proposed in contemporary models of intelligence structure (c.f. Carroll 1993; McGrew 2009). This model implies that individual differences in the different knowledge domains are driven by more general differences in declarative knowledge and that these individual differences are independent of differences in the administered response format. Although this model cannot be statistically tested against model C, because these models are not nested, model fit was sufficient for both models, so preferring the higher-order factor model due to its parsimony was justified. The higher-order factor captured substantial shares of variance ($p < .001$) and the factor saturation of the higher-order factor was high ($\omega_{Gc} = .89$; the reliability estimate was computed according to Brunner et al. 2012). Please note that we specified the same model with age as a predictor of Gc, and found that age was not a substantial predictor of Gc ($\beta = .13$, $p = .073$; $n = 300$; $\chi^2(61) = 112.44$, CFI = .958, RMSEA = .053, SRMR = .042).

Again, we added a defocusing count variable to the model ($\chi^2(61) = 80.17$, CFI = .984, RMSEA = .032, SRMR = .032) as a predictor of the higher-order Gc factor. The standardized regression weight was $\beta = .33$ ($SE = .07$). Defocusing, therefore, accounted for 11% of the

variance in the higher-order factor for crystallized intelligence. The remaining variance (i.e., individual differences after controlling for defocusing) was still significant ($p < .001$). The model is provided in Figure 3.

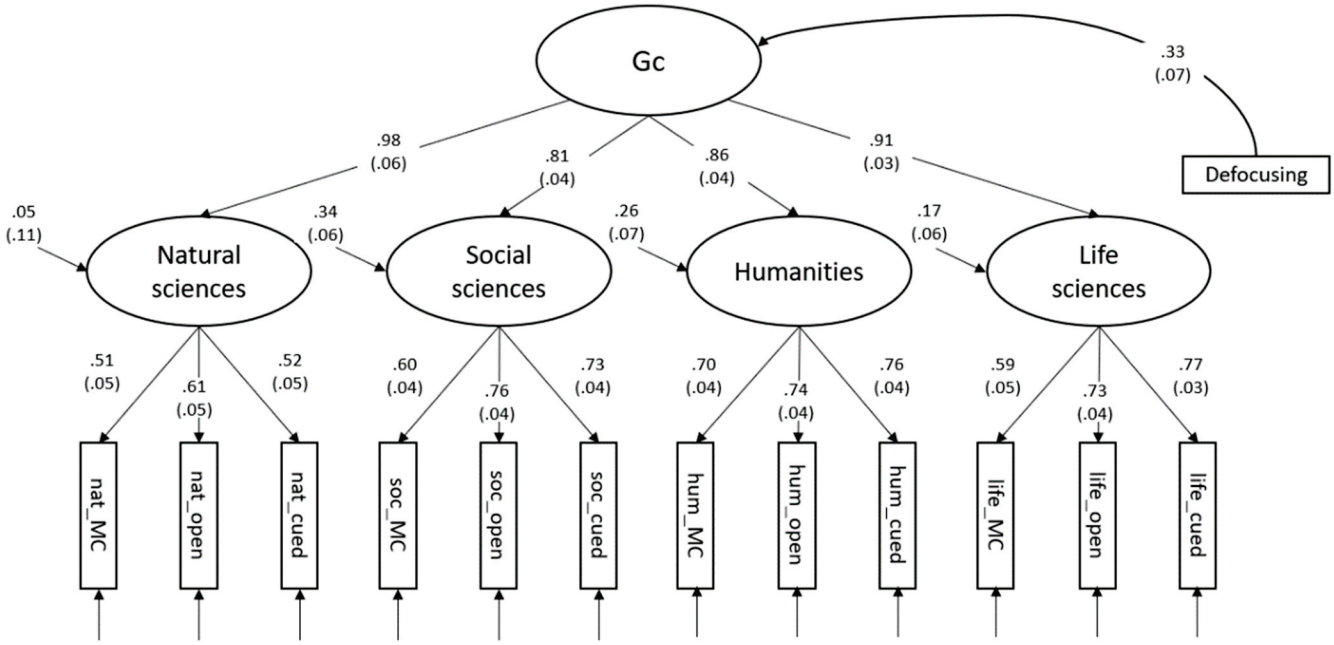


Figure 3. Higher-order model of the broad knowledge domains across response formats with defocusing as predictor. All parameters are standardized. $n = 300$; $\chi^2(61) = 80.17$, CFI = .984, RMSEA = .032, SRMR = .032. Standard errors are depicted in parentheses.

Eventually, we decided to further elaborate the higher-order factor model (model D of Table 3) by allowing for method variance due to the use of indicators of different response formats. We computed an additional model where we added two method factors to the model supposed to capture the joint variance of the open-ended and the cued open-ended response format. Theoretically, the nested method factors should explain individual differences in the indicators that are not already explained by the (domain-specific) knowledge factors. We chose the MC format as the reference method because this method is widely used across a range of applied fields. In sum, the model still fitted the data well ($n = 300$; $\chi^2(41) = 45.32$, CFI = .996, RMSEA = .019, SRMR = .025). The method factors were not able to account for any substantial shares of variance in the indicators and showed poor factor saturation ($\omega_{Cued} = .27$; $\omega_{Open} = .34$). We compared this model (method factors included) to the higher-order model (model D in Table 3) by a χ^2-difference test, which indicated that both models were not significantly different ($\Delta\chi^2(9, N = 300) = 16.65$, $p = .055$), and thus favoring the more parsimonious model without the proposed method factors. In sum, this means that it is sufficient to explain individual differences in the indicators with latent factors that are independent of the administered response format.

7. Discussion: Study 2

In study 2, we took advantage of the fact, that the 72 knowledge items from study 1 were homogenous regarding mean differences across response formats. A common supposition is that CR format items measure some cognitive processes such as remembering specific information from memory (i.e., knowledge) more readily than MC items (Martinez 1999; Rodriguez 2003). In fact, it was argued that CR format items even tap creative processes, because some items require new combinations of existing information (Schult and Sparfeldt 2018). If this were the case, the covariance between CR and MC tests should be reliably below unity. In study 2, we checked for the effects of manipulating response formats on the covariance structure of declarative fact knowledge items by means of confirmatory latent

variable modeling. As predicted, we confirmed and replicated that the mean structure is substantially affected by the response format. However, the rank-order of participants was unaffected by altering the response format. In other words, correlations between latent factors for MC and CR response formats reached unity. Individual differences in the knowledge tests were due to knowledge domains, which in turn could be subsumed below an overarching Gc factor. Response format turned out to be irrelevant to account for individual differences. From a multi-trait multi-method perspective (Campbell and Fiske 1959), method factors turned out to be irrelevant, whereas trait factors accounted for all of the individual differences.

8. General Discussion

We administered a broad fact knowledge test in two studies of the general population in three response formats: an MC format, a cued open-ended format, and an open-ended format. We tested how the empirical difficulties of the administered items were affected by response formats, and whether the rank-order of individuals taking the test changed across divergent response formats. Taken together, our results suggest that different response formats affect the mean structure of items, but the rank-orders of individuals remain the same.

8.1. Response Formats as Means to an End

Although the MC format is widely used (Chan and Kennedy 2002), it has been challenged in the literature many times (e.g., Becker and Johnston 1999; Hickson and Reed 2011; Krieg and Uyar 2001), and recommendations for using CR items as the superior item format are prevalent (e.g., Hickson and Reed 2011; Martinez 1999; Veloski et al. 1999).

From a measurement perspective, both the MC and the CR response formats should be understood as tools of measurement and therefore as specific means to an end. The question of whether both tools achieve their purpose equally well, in our case the assessment of declarative knowledge, has not yet been conclusively answered.

We identified three competing positions, from which different patterns of results can be derived: First, neither the mean structure nor the covariance structure of a test is affected by using different response formats. Second, only the mean-structure is affected by the response formats. Third, both the mean- and the covariance structure are affected by the response formats.

From an individual differences perspective, tests differing in their response formats would perform equally well, if the rank-order of individuals remains the same (Ackerman and Hambrick 2020). If the rank-order of individuals would change, this would indicate that the tests measure something different. If common assumptions that different response formats tap different forms of cognition (e.g., Buckles and Siegfried 2006) were true, abilities and methods for measuring them would be confounded.

Given the results, we conclude that measuring declarative knowledge with different response formats is possible without any loss of information regarding the rank-order of individuals. Although the mean structure was clearly affected by using different response formats, the rank-order of individuals did not change as a function of how we measured declarative knowledge. Although we cannot rule out that distinct cognitive processes are involved in answering fact knowledge questions in different response formats, as is predicted by the discontinuity hypothesis (Tulving and Watkins 1973), our data show that this is a concern that can be neglected if one is interested in individual differences. Clearly, correctly responding to fact knowledge items is a question of prior knowledge, but the possible underlying cognitive processes of identifying or retrieving a veridical solution to a knowledge question was irrelevant in accounting for individual differences in fact knowledge, which is in line with other recent studies on the subject (e.g., Schult and Sparfeldt 2018).

As all three administered response formats were indistinguishable psychometrically, some pragmatic considerations might lead to favoring one format over the other. From an

individual differences perspective, the costs and benefits of response format are mostly a methodological concern. For example, response formats could be seen as something that affects how easy tests can be instructed and scored, which relates to their being economically sound. In some instances, different response formats might be deemed different in terms of some reliability concepts. For example, we could ask how reliable a 20 items MC test is relative to an equally lengthy CR format test, but we could also ask how reliable a 20 min test session with MC items is relative to an equally long session with CR items. In terms of scoring, MC-based tests have a clear advantage. On the other side, battling with how guessing probability should be figured in has plagued generations of psychometricians.

8.2. Recognition, Recall, or What to Study Next

From a cognitive perspective, the distinction between recognition and recall is the most salient and striking feature that presumably distinguishes MC and CR item types (and in our case, the three administered response formats). Different response formats would then come along with different cognitive requirements. Often, these requirements are studied by comparing means in recognition and recall sessions with comparable materials, either within or between subjects. However, materials typically used in experiments that juxtapose recognition and recall are stimulus-sets over which experimental control can be exerted. For instance, intuitively it makes sense to ensure that stimulus-sets are completely new for all participants so that they can all go through a controlled learning phase prior to showing recognition or recall. This novelty can, for example, be ensured by creating artificial or meaningless stimulus-sets. This is not the case for a test battery of declarative knowledge items, because these items have to be meaningful by definition. Whether or not recognition and recall as measured by virtue of a declarative fact knowledge test with different response formats is psychometrically relevant (i.e., for explaining individual differences), is an empirical question. The present data suggest that response formats are not nearly as important as often suggested. A study allowing for latent variable modeling of declarative fact knowledge and broad retrieval abilities both measured with recognition and recall methods could show whether or not the results we report here extend to settings often used in experimental psychology.

Thus, broad retrieval abilities (Gr; Schneider and McGrew 2018) should be considered in further studies evaluating possible differences between MC format and CR format items. Broad retrieval abilities are understood as the rate and fluency with which individuals can selectively retrieve and produce information stored in long term-memory (Schneider and McGrew 2018). Established determinants of general retrieval ability are declarative knowledge (c.f., Hakstian and Cattell 1978; Jewsbury and Bowden 2017; Unsworth 2019) and working memory capacity (e.g., Rosen and Engle 1997; Unsworth 2019; Unsworth et al. 2011). As both prototypical retrieval tasks and open-ended declarative knowledge items require free production (i.e., responses to questions), a multivariate study investigating broad retrieval abilities, and declarative knowledge in MC and the here-used CR response formats could inform us further about whether different cognitive processes are tapped with different item formats for measuring declarative knowledge. Based on the present results, we predict that broad retrieval abilities are equally important for both types of item formats.

In addition to that, prior work shows that broad retrieval abilities can be explained through individual differences in working memory capacity (e.g., Rosen and Engle 1997). In turn, fluid intelligence is also strongly predicted by working memory capacity (Oberauer et al. 2005), due to the load put on working memory in prototypical reasoning tasks. The mean structures from the current studies suggest that different cognitive processes in responding to items might be tapped. Whereas the MC format urges test takers to recognize the correct response by comparing the provided response alternatives to what the test takers might have stored in memory, the open-ended formats lead them to exhaustively and proactively search for the correct response. Arguably, the latter process puts more demand on working memory, as the individuals have to bear in mind the responses they have already discarded. This makes sense and is in line with previous literature showing

that WMC is needed for controlled searches in long-term memory (e.g., Unsworth and Brewer 2009). In contrast, the MC format already provides some particular response alternatives, and these do not have to be kept in mind, because they are accessible to the respondents without any cognitive effort. Therefore, in contrast to CR formats, MC formats have been argued to have the smallest risk of being confounded by fluid intelligence (Schipolowski et al. 2014), because they impose lower demands on working memory.

Please note that the irrelevance of the response format for explaining the covariance structure, as observed in the present study, has an important implication. If performance in MC response format items, which supposedly tap recognition, and open-ended response format tests, which arguably tap recall, are perfectly correlated, they cannot be differentially related with other variables. For instance, recall tests might be argued to stress working memory more strongly by virtue of requiring subjects to maintain a list of candidate responses they discard as erroneous. However, the present results are at odds with such differential relations. Therefore, it might be argued that the here-presented results deviate from previous findings stressing that recognition and recall can be psychometrically distinguished (e.g., Unsworth and Brewer 2009; Unsworth 2019). However, it is important to note that differences in operationalizing vague terms such as "recognition" and "recall" should be considered when interpreting the present data. Studies in which recall (also termed familiarity) and recognition (also termed recollection) can be psychometrically distinguished (e.g., Unsworth and Brewer 2009) rely on tasks measuring memory for previously learned lists (e.g., word lists). The present study presupposes an understanding of the terms "recognition" and "recall" in the context of fact knowledge, which intertwines recognition and recall with the general knowledge that is either existent in long-term memory or not. Caution should thus be exerted in directly comparing results based on different study paradigms.

In fact, the term "recognition" suggests that the solution to MC items is stored in a verbatim fashion in LTM and that the distractors do not compete in being recognized. Both assumptions are likely to be wrong. It is implausible that question-response pairs are deposited in LTM waiting to be triggered by a recognition question one day. Often enough responses to fact knowledge questions require elaborate questions and also lengthy responses. The term recognition, therefore, downplays what is performed in such MC questions. In some instances, it might be recognition of a response that is stored in a verbatim fashion in LTM. In other instances, solving MC items requires weighting response options in terms of plausibility or probability prior to responding. In still other instances, the correct solution might be inferred by eliminating distractors. In yet other instances, a lucky punch by pure guessing might be what is observed. Needless to say, these and many different paths to a response are unlikely to be fixed for an item. Whereas some persons might simply retrieve a solution, others might be guessing, and still others might weigh the evidence of each response option and the roles of these participants might change in the next MC question.

This might also be an effect of item content. Future research should investigate whether different items might be more or less associated with semantic or episodic memory and if this predicts how difficult they are to solve when accounting for individual learning histories. Fact knowledge questions should refer to knowledge that persons can actually acquire during their lives (Wilhelm and Schroeders 20219, but as this is an idiosyncratic aspect, differences between response formats with respect to which kind of information the items tap might be considered. Such a study could be conducted via an experience-sampling method with several groups, where, across time, the different groups are exposed to different learning environments tapping different topics. After an a priori determined time of several weeks with relatively constant learning histories, the newly acquired knowledge of the participants could be tested. To then determine whether the individual learning history of the participants plays a role in responding to certain items from certain knowledge domains in competing response formats, all participants will be not only tested

with regard to their, in light of the study, personal learning history, but also on all topics from the other experimental groups.

In addition to that, future research could test whether the correlation between open-ended and MC item formats might hinge upon the age of the test takers. If the association between these response formats changed as a function of age, this would indicate that the cognitive processes underlying the response behavior might be subject to change, too; which might correspond to a decline in retrieval ability in advanced age.

8.3. Limitations

The present studies feature important strengths such as the administration of a broad test battery of declarative knowledge to two independent samples of the general population, and the use of suitable modeling approaches (e.g., logistic mixed regression and latent variable models). A key limitation of the present research designs is that both studies were conducted online and unproctored. These limitations should introduce noise into measurement which should make it more difficult to find reliable associations. However, the key result we report is a perfect correlation of latent variables for recognition (i.e., MC items) and recall (i.e., open-ended and cued open-ended items). We made effort to remove invalid observations, for example by using a control mechanism to observe whether participants changed tabs during the test sessions (Diedenhofen and Musch 2017). Nevertheless, future research should replicate the present results in a proctored laboratory test setting in an effort to contribute to replicable and cumulative research (Flake and Fried 2020). Lastly, it should be noted that our results, especially regarding the construct validity, do not necessarily transfer to measurements of other cognitive abilities for which different response formats can be administered (e.g., reasoning ability as measured with matrix construction tests; c.f., Becker et al. 2015, 2016).

9. Conclusions

The effects of response formats on tests of declarative fact knowledge showed expected mean differences but no distinction on the level of latent variables. If they have the option to choose, researchers and practitioners should use MC-format items, as this response format is easier to administer and cheaper to score, and declarative fact knowledge does not change as a function of the response format. Rather, individual differences in declarative fact knowledge are best expressed as a high-order factor model with broad domain factors below an overarching Gc factor.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jintelligence10040102/s1.

Author Contributions: Conceptualization, B.G. and O.W.; Data curation, B.G., C.S., and M.S.; Formal analysis, B.G.; Investigation, B.G.; Methodology, B.G. and O.W.; Project administration, B.G.; Resources, O.W.; Software, B.G., C.S., and M.S.; Supervision, O.W.; Validation, B.G. and O.W.; Visualization, B.G.; Writing—original draft, B.G.; Writing—review & editing, B.G., O.W., C.S. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: The work described in this study was carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). All participants provided informed consent by starting the online study.

Data Availability Statement: To make all analyses reproducible, we provide all materials necessary, including data and analysis scripts, in an online repository: https://osf.io/pse3w/.

Acknowledgments: We thank Amina Rajakumar for additional scoring of the items and Diana Steger and related colleagues for the initial compilation of the item pool.

Conflicts of Interest: The authors declare no conflict of interest.

References

Abu-Zaid, Ahmed, and Tehreem A. Khan. 2013. Assessing Declarative and Procedural Knowledge Using Multiple-Choice Questions. *Medical Education Online* 18: 21132. [CrossRef] [PubMed]

Ackerman, Phillip L. 1996. A Theory of Adult Intellectual Development: Process, Personality, Interests, and Knowledge. *Intelligence* 22: 227–57. [CrossRef]

Ackerman, Phillip 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. [CrossRef] [PubMed]

Ackerman, Phillip L., and David Z. Hambrick. 2020. A Primer on Assessing Intelligence in Laboratory Studies. *Intelligence* 80: 101440. [CrossRef]

Amthauer, Rudolf, Burkhard Brocke, Detlev Liepmann, and André Beauducel. 2001. *Intelligenz-Struktur-Test 2000 R Manual [Manual of the Intelligence Structure Test 2000 R]*. Göttingen: Hogrefe.

Anderson, John R., and Gordon H. Bower. 1972. Recognition and Retrieval Processes in Free Recall. *Psychological Review* 79: 97–123. [CrossRef]

Beauducel, André, and Martin Kersting. 2002. Fluid and Crystallized Intelligence and the Berlin Model of Intelligence Structure (BIS). *European Journal of Psychological Assessment* 18: 97–112. [CrossRef]

Beauducel, Andre, and Philipp Yorck Herzberg. 2006. On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal* 13: 186–203. [CrossRef]

Becker, Nicolas, Florian Schmitz, Anke Falk, Jasmin Feldbrügge, Daniel Recktenwald, Oliver Wilhelm, Franzis Preckel, and Frank Spinath. 2016. Preventing Response Elimination Strategies Improves the Convergent Validity of Figural Matrices. *Journal of Intelligence* 4: 2. [CrossRef]

Becker, Nicolas, Franzis Preckel, Julia Karbach, Nathalie Raffel, and Frank M. Spinath. 2015. Die Matrizenkonstruktionsaufgabe: Validierung eines distraktorfreien Aufgabenformats zur Vorgabe figuraler Matrizen. *Diagnostica* 61: 22–33. [CrossRef]

Becker, William E., and Carol Johnston. 1999. The Relationship between Multiple Choice and Essay Response Questions in Assessing Economics Understanding. *Economic Record* 75: 348–57. [CrossRef]

Bentler, Peter M. 1990. Comparative Fit Indexes in Structural Models. *Psychological Bulletin* 107: 238–46. [CrossRef] [PubMed]

Browne, Michael W., and Robert Cudeck. 1992. Alternative Ways of Assessing Model Fit. *Sociological Methods & Research* 21: 230–58.

Brunner, Martin, Gabriel Nagy, and Oliver Wilhelm. 2012. A Tutorial on Hierarchically Structured Constructs. *Journal of Personality* 80: 796–846. [CrossRef]

Buckles, Stephen, and John J. Siegfried. 2006. Using Multiple-Choice Questions to Evaluate In-Depth Learning of Economics. *The Journal of Economic Education* 37: 48–57. [CrossRef]

Campbell, Donald T., and Donald W. Fiske. 1959. Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin* 56: 81–105. [CrossRef] [PubMed]

Carroll, John B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*, 1st ed. Cambridge: Cambridge University Press. [CrossRef]

Cattell, Raymond B. 1957. *Personality and Motivation Structure and Measurement*. Chicago: World Book.

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

Cattell, Raymond B. 1987. *Intelligence: Its Structure, Growth, and Action*. Advances in Psychology 35. Amsterdam and New York: Elsevier Science Pub. Co.

Chan, Nixon, and Peter E. Kennedy. 2002. Are Multiple-Choice Exams Easier for Economics Students? A Comparison of Multiple-Choice and 'Equivalent' Constructed-Response Exam Questions. *Southern Economic Journal* 68: 957–71. [CrossRef]

Chapelle, Carol A. 1999. Construct Definition and Validity Inquiry in SLA Research. In *Interfaces between Second Language Acquisition and Language Testing Research*, 1st ed. Edited by Lyle F. Bachman and Andrew D. Cohen. Cambridge: Cambridge University Press, pp. 32–70. [CrossRef]

Chittooran, Mary M., and Dorothy D. Miles. 2001. *Test-Taking Skills for Multiple-Choice Formats: Implications for School Psychologists*. Washington, DC: Education Resources Information Center, p. 23.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20: 37–46. [CrossRef]

Cohen, Jacob. 1969. *Statistical Power Analysis for the Behavioral Sciences*. San Diego: Academic Press.

Cole, David A., Corinne E. Perkins, and Rachel L. Zelkowitz. 2016. Impact of Homogeneous and Heterogeneous Parceling Strategies When Latent Variables Represent Multidimensional Constructs. *Psychological Methods* 21: 164–74. [CrossRef]

Coleman, Chris, Jennifer Lindstrom, Jason Nelson, William Lindstrom, and K. Noël Gregg. 2010. Passageless Comprehension on the Nelson-Denny Reading Test: Well Above Chance for University Students. *Journal of Learning Disabilities* 43: 244–49. [CrossRef]

Daneman, Meredyth, and Brenda Hannon. 2001. Using Working Memory Theory to Investigate the Construct Validity of Multiple-Choice Reading Comprehension Tests Such as the SAT. *Journal of Experimental Psychology: General* 130: 208–23. [CrossRef]

Diedenhofen, Birk, and Jochen Musch. 2017. PageFocus: Using Paradata to Detect and Prevent Cheating on Online Achievement Tests. *Behavior Research Methods* 49: 1444–59. [CrossRef] [PubMed]

Downing, Steven M., and Thomas M. Haladyna, eds. 2006. *Handbook of Test Development*. Mahwah: L. Erlbaum.

Enders, Craig K., ed. 2010. *Applied Missing Data Analysis*. New York: Guilford Press.

Fazio, Lisa K., Pooja K. Agarwal, Elizabeth J. Marsh, and Henry L. Roediger. 2010. Memorial Consequences of Multiple-Choice Testing on Immediate and Delayed Tests. *Memory & Cognition* 38: 407–18. [CrossRef]
Flake, Jessica Kay, and Eiko I. Fried. 2020. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science* 3: 456–65. [CrossRef]
Fowler, Bill, and Barry M. Kroll. 1978. Verbal Skills as Factors in the Passageless Validation of Reading Comprehension Tests. *Perceptual and Motor Skills* 47: 335–38. [CrossRef]
Gillund, Gary, and Richard M. Shiffrin. 1984. A Retrieval Model for Both Recognition and Recall. *Psychological Review* 91: 67. [CrossRef]
Haist, Frank, Arthur P. Shimamura, and Larry R. Squire. 1992. On the Relationship Between Recall and Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18: 691–702. [CrossRef] [PubMed]
Hakstian, A. Ralph, and Raymond B. Cattell. 1978. Higher-Stratum Ability Structures on a Basis of Twenty Primary Abilities. *Journal of Educational Psychology* 70: 657–69. [CrossRef]
Hancock, Gregory R. 1994. Cognitive Complexity and the Comparability of Multiple-Choice and Constructed-Response Test Formats. *The Journal of Experimental Education* 62: 143–58. [CrossRef]
Harke, Douglas J., J. Dudley Herron, and Ralph W. Leffler. 1972. Comparison of a Randomized Multiple Choice Format with a Written One-Hour Physics Problem Test. *Science Education* 56: 563–65. [CrossRef]
Hartung, Johanna, Selina Weiss, and Oliver Wilhelm. 2017. Individual Differences in Performance on Comprehension and Knowledge Tests with and without Passages and Questions. *Learning and Individual Differences* 56: 143–50. [CrossRef]
Hickson, Stephen, and Bob Reed. 2011. More Evidence on the Use of Constructed-Response Questions in Principles of Economics Classes. *International Review of Economics Education* 10: 28–49. [CrossRef]
Hohensinn, Christine, and Klaus D. Kubinger. 2011. Applying Item Response Theory Methods to Examine the Impact of Different Response Formats. *Educational and Psychological Measurement* 71: 732–46. [CrossRef]
Horn, John L. 1965. Fluid and Crystallized Intelligence: A Factor Analytic Study of the Structure among Primary Mental Abilities. Ph.D. dissertation, University of Illinois, Urbana, IL, USA.
Horn, John L. 1966. Some Characteristics of Classroom Examinations. *Journal of Educational Measurement* 3: 293–95. [CrossRef]
Hu, Li-tze, and Peter M. Bentler. 1999. Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6: 1–55. [CrossRef]
Jewsbury, Paul A., and Stephen C. Bowden. 2017. Construct Validity of Fluency and Implications for the Factorial Structure of Memory. *Journal of Psychoeducational Assessment* 35: 460–81. [CrossRef]
Katz, Stuart, A. Boyd Blackburn, and Gary J. Lautenschlager. 1991. Answering Reading Comprehension Items without Passages on the SAT When Items Are Quasi-Randomized. *Educational and Psychological Measurement* 51: 747–54. [CrossRef]
Katz, Stuart, Gary J. Lautenschlager, A. Boyd Blackburn, and Felicia H. Harris. 1990. Answering Reading Comprehension Items without Passages on the SAT. *Psychological Science* 1: 122–27. [CrossRef]
Kennedy, Peter, and William B. Walstad. 1997. Combining Multiple-Choice and Constructed-Response Test Scores: An Economist's View. *Applied Measurement in Education* 10: 359–75. [CrossRef]
Kesselman-Turkel, Judi, and Franklynn Peterson. 2004. *Test-Taking Strategies*. Madison: University of Wisconsin Press.
Krathwohl, David R. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice* 41: 212–18. [CrossRef]
Krieg, Randall G., and Bulent Uyar. 2001. Student Performance in Business and Economics Statistics: Does Exam Structure Matter? *Journal of Economics and Finance* 25: 229–41. [CrossRef]
Li, Shu-Chen, Ulman Lindenberger, Bernhard Hommel, Gisa Aschersleben, Wolfgang Prinz, and Paul B. Baltes. 2004. Transformations in the Couplings among Intellectual Abilities and Constituent Cognitive Processes across the Life Span. *Psychological Science* 15: 155–63. [CrossRef] [PubMed]
Lindner, Marlit Annalena, Jörn R. Sparfeldt, Olaf Köller, Josef Lukas, and Detlev Leutner. 2021. Ein Plädoyer zur Qualitätssicherung schriftlicher Prüfungen im Psychologiestudium. *Psychologische Rundschau* 72: 93–105. [CrossRef]
Little, Todd D., William A. Cunningham, Golan Shahar, and Keith F. Widaman. 2002. To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits. *Structural Equation Modeling: A Multidisciplinary Journal* 9: 151–73. [CrossRef]
Lukhele, Robert, David Thissen, and Howard Wainer. 1994. On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests. *Journal of Educational Measurement* 31: 234–50. [CrossRef]
Lynn, Richard, and Paul Irwing. 2002. Sex Differences in General Knowledge, Semantic Memory and Reasoning Ability. *British Journal of Psychology* 93: 545–56. [CrossRef]
Lynn, Richard, Paul Irwing, and Thomas Cammock. 2001. Sex Differences in General Knowledge. *Intelligence* 30: 27–39. [CrossRef]
Martinez, Michael E. 1999. Cognition and the Question of Test Item Format. *Educational Psychologist* 34: 207–18. [CrossRef]
McDonald, Roderick P. 1999. *Test Theory: A Unified Treatment*. Hillsdale: Erlbaum.
McGrew, Kevin S. 2005. The Cattell-Horn-Carroll Theory of Cognitive Abilities: Past, Present, and Future. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. New York: The Guilford Press, pp. 136–81.
McGrew, Kevin 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. [CrossRef]
Millman, Jason, Carol H. Bishop, and Robert Ebel. 1965. An Analysis of Test-Wiseness. *Educational and Psychological Measurement* 25: 707–26. [CrossRef]

Mullis, Ina V. S., Michael O. Martin, and Pierre Foy. 2008. *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Mullis, Ina V. S., Michael O. Martin, Ann M. Kennedy, and Pierre Foy. 2007. *IEA's Progress in International Reading Literacy Study in Primary School in 40 Countries*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Nakagawa, Shinichi, Paul C. D. Johnson, and Holger Schielzeth. 2017. The Coefficient of Determination R2 and Intra-Class Correlation Coefficient from Generalized Linear Mixed-Effects Models Revisited and Expanded. *Journal of the Royal Society* 14: 11.

Oberauer, Klaus, Ralf Schulze, Oliver Wilhelm, and Heinz-Martin Süß. 2005. Working Memory and Intelligence—Their Correlation and Their Relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin* 131: 61–65. [CrossRef] [PubMed]

Pornprasertmanit, Sunthud, Patrick Miller, Alexander Schoemann, and Terrence D. Jorgensen. 2021. Simsem: SIMulated Structural Equation Modeling. R Package. Available online: https://CRAN.Rproject.org/package=simsem (accessed on 1 January 2022).

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 1 January 2022).

Raykov, Tenko, and George A. Marcoulides. 2011. Classical Item Analysis Using Latent Variable Modeling: A Note on a Direct Evaluation Procedure. *Structural Equation Modeling: A Multidisciplinary Journal* 18: 315–24. [CrossRef]

Rodriguez, Michael C. 2003. Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement* 40: 163–84. [CrossRef]

Rosen, Virginia M., and Randall W. Engle. 1997. The Role of Working Memory Capacity in Retrieval. *Journal of Experimental Psychology: General* 126: 211–27. [CrossRef]

Rosseel, Yves. 2012. Lavaan: An R Package for Structural Equation Moeling. *Journal of Statistical Software* 48: 1–36. [CrossRef]

Rost, Detlef H., and Jörn R. Sparfeldt. 2007. Leseverständnis ohne Lesen?: Zur Konstruktvalidität von multiple-choice-Leseverständnistestaufgaben. *Zeitschrift für Pädagogische Psychologie* 21: 305–14. [CrossRef]

Sabers, Darrell. 1975. *Test-Taking Skills*. Tucson: The University of Arizona.

Sam, Amir H., Emilia Peleva, Chee Yeen Fung, Nicki Cohen, Emyr W. Benbow, and Karim Meeran. 2019. Very Short Answer Questions: A Novel Approach To Summative Assessments In Pathology. *Advances in Medical Education and Practice* 10: 943–48. [CrossRef]

Sam, Amir H., Samantha M. Field, Carlos F. Collares, Colin Melville, Joanne Harris, and Karim Meeran. 2018. Very-short-answer Questions: Reliability, Discrimination and Acceptability. *Medical Education* 52: 447–55. [CrossRef]

Schafer, Joseph L., and John W. Graham. 2002. Missing Data: Our View of the State of the Art. *Psychological Methods* 7: 147–77. [CrossRef]

Schipolowski, Stefan, Oliver Wilhelm, and Ulrich Schroeders. 2014. On the Nature of Crystallized Intelligence: The Relationship between Verbal Ability and Factual Knowledge. *Intelligence* 46: 156–68. [CrossRef]

Schneider, W. Joel, and Kevin S. McGrew. 2018. The Cattell–Horn–Carroll Theory of Cognitive Abilities. In *Contemporary Intellectual Assessment: Theories, Tests and Issues*. New York: Guilford Press, p. 91.

Schroeders, Ulrich, Oliver Wilhelm, and Gabriel Olaru. 2016. The Influence of Item Sampling on Sex Differences in Knowledge Tests. *Intelligence* 58: 22–32. [CrossRef]

Schroeders, Ulrich, Stefan Schipolowski, and Oliver Wilhelm. 2020. *Berliner Test Zur Erfassung Fluider Und Kristalliner Intelligenz Für Die 5. Bis 7. Klasse (BEFKI 5-7)*. Göttingen: Hogrefe Verlag.

Schult, Johannes, and Jörn R. Sparfeldt. 2018. Reliability and Validity of PIRLS and TIMSS: Does the Response Format Matter? *European Journal of Psychological Assessment* 34: 258–69. [CrossRef]

Scouller, Karen. 1998. The Influence of Assessment Method on Students' Learning Approaches: Multiple Choice Question Examination versus Assignment Essay. *Higher Education* 35: 453–72. [CrossRef]

Scully, Darina. 2017. Constructing Multiple-Choice Items to Measure Higher-Order Thinking. *Practical Assessment, Research, and Evaluation Practical Assessment, Research, and Evaluation* 22: 4. [CrossRef]

Sparfeldt, Jörn R., Rumena Kimmel, Lena Löwenkamp, Antje Steingräber, and Detlef H. Rost. 2012. Not Read, but Nevertheless Solved? Three Experiments on PIRLS Multiple Choice Reading Comprehension Test Items. *Educational Assessment* 17: 214–32. [CrossRef]

Steger, Diana, Ulrich Schroeders, and Oliver Wilhelm. 2019. On the Dimensionality of Crystallized Intelligence: A Smartphone-Based Assessment. *Intelligence* 72: 76–85. [CrossRef]

Steger, Diana, Ulrich Schroeders, and Oliver Wilhelm. 2020. Caught in the Act: Predicting Cheating in Unproctored Knowledge Assessment. *Assessment* 28: 1004–17. [CrossRef] [PubMed]

Thissen, David, Howard Wainer, and Xiang-Bo Wang. 1994. Are Tests Comprising Both Multiple-Choice and Free-Response Items Necessarily Less Unidimensional Than Multiple-Choice Tests? An Analysis of Two Tests. *Journal of Educational Measurement* 31: 113–23. [CrossRef]

Traub, Ross E. 1993. On the Equivalence of the Traits Assessed by Multiple-Choice and Constructed-Response Tests. In *Construction Versus Choice in Cognitive Measurement. Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. New York: Routledge, pp. 29–44.

Traub, Ross E., and Charles W. Fisher. 1977. On the Equivalence of Constructed- Response and Multiple-Choice Tests. *Applied Psychological Measurement* 1: 355–69. [CrossRef]

Tulving, Endel, and Michael J. Watkins. 1973. Continuity between Recall and Recognition. *The American Journal of Psychology* 86: 739. [CrossRef]

Unsworth, Nash, and Gene A. Brewer. 2009. Examining the Relationships among Item Recognition, Source Recognition, and Recall from an Individual Differences Perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35: 1578–85. [CrossRef]

Unsworth, Nash, Gregory J. Spillers, and Gene A. Brewer. 2011. Variation in Verbal Fluency: A Latent Variable Analysis of Clustering, Switching, and Overall Performance. *The Quarterly Journal of Experimental Psychology* 64: 447–66. [CrossRef] [PubMed]

Unsworth, Nash. 2019. Individual Differences in Long-Term Memory. *Psychological Bulletin* 145: 79–139. [CrossRef] [PubMed]

Veloski, J. Jon, Howard K. Rabinowitz, Mary R. Robeson, and Paul R. Young. 1999. Patients Don't Present with Five Choices: An Alternative to Mulitple-Choice Tests in Assessing Physicians' Competence. *Academic Medicine* 74: 539–46. [CrossRef] [PubMed]

Vernon, Philip E. 1962. The Determinants of Reading Comprehension. *Educational and Psychological Measurement* 22: 269–86. [CrossRef]

von Stumm, Sophie, and Phillip L. Ackerman. 2013. Investment and Intellect: A Review and Meta-Analysis. *Psychological Bulletin* 139: 841–69. [CrossRef]

Walstad, William B. 2001. Improving Assessment in University Economics. *The Journal of Economic Education* 32: 281–94. [CrossRef]

Ward, William C. 1982. A Comparison of Free-Response and Multiple-Choice Forms of Verbal Aptitude Tests. *Applied Psychological Measurement* 6: 1–11. [CrossRef]

Watrin, Luc, Ulrich Schroeders, and Oliver Wilhelm. 2022. Structural Invariance of Declarative Knowledge across the Adult Lifespan. *Psychology and Aging* 37: 283–97. [CrossRef]

Wilhelm, Oliver, and Patrick Kyllonen. 2021. To Predict the Future, Consider the Past: Revisiting Carroll (1993) as a Guide to the Future of Intelligence Research. *Intelligence* 89: 101585. [CrossRef]

Wilhelm, Oliver, and Ulrich Schroeders. 2019. Intelligence. In *The Psychology of Human Thought: An Introduction*. Edited by Robert Sternberg and Joachim Funke. Heidelberg: Heidelberg University Publishing, pp. 257–77.

V Manuscript IV

Driving Factors of Individual Differences in Broad Retrieval Ability: Gr *Is More Than the Sum of Its Parts*

The study presented in this manuscript was initially conceptualized by all co-authors. Dr. Selina Weiss and myself programmed and piloted the test battery. The organization and data collection were supervised by Dr. Selina Weiss and myself. Data analysis regarding the manuscript at hand was mainly conducted by myself under the supervision of all co-authors. I have completed a literature search and drafted the original manuscript. All co-authors reviewed and revised the manuscript under my supervision as first author. The manuscript is currently in preparation for submission in a peer-reviewed journal, but has already been published as a preprint in an open science repository. It has not yet been peer-reviewed.

Goecke, B., Weiss, S., & Wilhelm, O. (2023). *Driving Factors of Individual Differences in Broad Retrieval Ability: Gr Is More Than the Sum of Its Parts* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/zy7mp

Driving Factors of Individual Differences in Broad Retrieval Ability: *Gr* Is More Than the Sum of Its Parts

Goecke, B., Weiss, S., & Wilhelm, O.

Ulm University, Ulm, Germany

Benjamin Goecke https://orcid.org/0000-0002-3050-1848

Selina Weiss https://orcid.org/0000-0001-8796-2154

Oliver Wilhelm https://orcid.org/0000-0001-7980-1166

Author Note

Correspondence concerning this article should be addressed to Benjamin Goecke, Institute for Psychology and Pedagogy, Ulm University, Albert-Einstein-Allee 47, 89081 Ulm, Germany. Telephone: +49-(0)731/50 31146. Fax: +49-(0)731/50 31149. E-mail: Benjamin.Goecke95@gmail.com

Acknowledgements

We thank Amina Rajakumar, Jasmin Thelen, and Tim Trautwein for their support in data collection.

This manuscript is a preprint version of a manuscript (i.e., before peer-review).

Update: 17-04-2023

© This manuscript is made available under the CC-BY 4.0 license by the Authors.

Abstract

Broad retrieval ability (Gr) posits an essential factor of human cognitive abilities. Previous literature indicates Gr is best modeled as a higher-order factor model with lower-level factors such as ideational, word, expressional, or figural fluency. However, the dimensionality of Gr is not well studied. Further, it remains unclear whether specific retrieval affordances like differing retrieval time periods (e.g., short vs. long) can be psychometrically separated from the general ability. Such a distinction would imply differential associations between specific retrieval, general retrieval, and other cognitive abilities, which, in turn, depict a vital part of explanatory models of individual differences in Gr. To test these assumptions, we conducted a multivariate study (N = 331) and evaluated competing latent variable measurement models for a variety of Gr tests. We then regressed the best measurement model onto working memory capacity, secondary memory, mental speed, and crystallized intelligence in order to evaluate the distinctiveness of Gr. Our results suggest that no specific retrieval affordances with regards to time periods can be distinguished. A higher-order model, depicting a second-order Gr factor with lower-order factors underneath (ideational, word, expressional, and figural fluency) fitted the data best, extending previous literature by increasing construct coverage. All covariates show incremental predictive validity, beyond their shared variances. In sum, our results (a) endorse a perspective on Gr as a strong and discriminant factor of cognitive abilities that is not affected by time constraints, and (b) show that Gr is more than a linear combination of its parts.

Keywords: *broad retrieval ability, Gr, fluency, individual differences, cognitive abilities*

Driving Factors of Individual Differences in Broad Retrieval Abilities: *Gr* Is More Than the Sum of Its Parts

Introduction

Broad retrieval ability (Gr) refers to an individual's ability to retrieve and (re)produce verbal and nonverbal information from long-term memory (Schneider & McGrew, 2018). This ability is measured by tasks that assess *fluency*, that is, the number of instances of a given category (e.g., animals) test takers can retrieve in a limited amount of time (e.g., 60 seconds). Retrieval ability is relevant because it is related to age-associated decline of cognitive functioning (e.g., Mayr & Kliegl, 2000; Troyer et al., 1997), deficits in neuropsychological functioning (e.g., Metternich et al., 2014) and creativity (e.g., Silvia et al., 2013). High retrieval ability has been associated with better mental speed (Miroshnik et al., 2023), higher working memory capacity (WMC; Rosen & Engle, 1997), larger secondary memory (i.e., long-term memory; Unsworth, 2019), and greater crystallized intelligence (Gc; Jewsbury & Bowden, 2017), all of which were discussed as important for explaining individual differences in Gr.

Psychometric models of Gr suggest a higher-order factor model as an adequate representation of Gr (Carroll, 1993; Schneider & McGrew, 2018), but relatively sparse efforts have been made to test this assumption by means of confirmatory factor analysis (for exceptions, see Jewsbury & Bowden, 2017 and Silvia et al., 2013). In contrast to that, driving factors of individual differences in Gr are better understood (e.g., Rosen & Engle, 1997; Unsworth et al., 2011), but multivariate research combining the establishment of a broad measurement model based on several narrow abilities of Gr with an examination of various candidate predictors in a latent variable approach is still missing. With the current study, we aimed to close this research gap.

Dimensionality of Gr

Gr has been studied as a mechanical ability (i.e., an ability of basic information processing; Baltes et al., 1999) in various taxonomic models of intelligence (e.g., three-stratum-model, Carroll, 1993; Carroll-Horn-Cattel-model, Schneider & McGrew, 2018). Several narrow abilities are subsumed underneath a higher-order factor, mainly representing verbal instances of Gr (i.e., ideational fluency, expressional fluency,

associational fluency, sensitivity to problems/alternative solution fluency, originality/creativity, speed of lexical access, naming facility, and word fluency (c.f., Schneider & McGrew, 2018)). Although non-verbal stimulus domains are also present in this taxonomic approach (e.g., figural fluency/flexibility), the roots of studying Gr lie in studying verbal fluency tests like word-association tests (e.g., Thurstone, 1938). This tradition has endured to this day (Unsworth, 2019) and should be viewed as a limitation in the study of Gr.

In the few previous studies investigating the dimensionality of Gr, Gr fitted the data best as a higher-order factor with six underlying first-order factors, including word, ideational, associational, and letter fluency, and associative flexibility and dissociative ability (Silvia et al., 2013), but the authors did not report further competing measurement models of Gr. Other work (Jewsbury & Bowden, 2017) evaluated the structure of narrow Gr factors according to the CHC model (Schneider & McGrew, 2018) by re-analyzing data from Carroll (1993). For each data set, the authors tested various competing measurement models of correlated group factors. Models included tests of associative, expressional, ideational, and word fluency, and naming facility and verbal closure, but also covariates like fluid and crystallized intelligence and mental speed. Concluding, the authors did not find evidence for narrow abilities as proposed by the CHC taxonomy but instead argued for an empirical distinction between semantic and orthographic fluency tests (Robinson et al., 2012). Taken together, these results stand against the suggestion of a higher-order Gr factor.

Furthermore, the above-mentioned studies sole focus on verbal indicators of Gr is problematic. Frequent omissions of figural fluency tasks in the literature (for an exception, see e.g., Vannorsdall et al., 2012) prevent understanding Gr "in its broadest possible sense" (Carroll, 1993, p. 394). We argue that Gr should best be measured by an adequate coverage of indicators, including figural Gr tests. To understand a cognitive ability construct holistically, it is essential to consider various instantiations of the tasks (Campbell & Fiske, 1959). Only this can ensure that Gr is grasped adequately with respect to its definition, mentioning verbal and non-verbal information in long-term memory (c.f., Schneider & McGrew, 2018).

Yet another task-related characteristic that might affect a measurement model of Gr pertains to the duration of typical retrieval tasks in the verbal domain. Prototypical verbal retrieval tasks are designed to take 60 seconds, and only a few studies apply tasks that take longer (e.g., 5 minutes, Schelble et al., 2012; Unsworth et al., 2013; 15 minutes, Rosen & Engle, 1997). To our knowledge, no study regarding individual differences in Gr has examined whether the differences in the pre-defined time frames in retrieval tasks add to individual differences (i.e., rank-order differences) in Gr performance. If this was the case, this would have serious practical implications like the need to re-evaluate previous empirical results.

This poses an interesting research question: do time frames of Gr tasks play a role for individual differences, and if yes, why? If performance in longer task times as compared to performance in shorter task times in retrieval tasks was differentially related to other cognitive abilities like WMC, differing retrieval demands (i.e., short vs. long) should be psychometrically separable in a measurement model of Gr and it remains an open question whether such results generalize across other retrieval tasks. Moreover, it remains an open question whether differences between short and long periods of retrieval time capture different aspects of individual differences in Gr in the first place.

Cognitive Processes Underlying Gr

In order to better understand Gr, it is not only important to establish a sound measurement model, but it is also helpful to examine competing models which explain individual differences in retrieval ability. Such models can help in identifying the mechanisms that enable cognitive processes involved in retrieval ability (c.f., Frischkorn et al., 2022).

Classical two-stage models (e.g., Bousfield & Sedgewick, 1944; Gruenewald & Lockhead, 1980; Wixted & Rohrer, 1994) articulate the idea that retrieval performance is achieved by repeating the scheme of an initial search for overall categories, which is then followed by sampling single representatives from these categories. For instance, test takers demanded to retrieve animals might first start with finding a more general category like mammals or farm animals and only then start to retrieve single animals from these respective categories. This process was thought to be repeated until either the time of the task ends or until all categories have been exhausted; which, under usual time constraints

(i.e., 60 seconds), is arguably seldom the case. Generally, two-stage models are well-supported (c.f., Unsworth et al., 2011) and are able to explain frequently observed patterns in verbal fluency data (e.g., representatives from one subcategory are retrieved rapidly after one another, but there are time-related breaks between subcategories). However, these models fail to explain empirical findings such as the tendency for individuals to avoid returning to already-retrieved responses (Azuma, 2004; Rosen & Engle, 1997). Thus, ideas revolving around the cognitive processes underlying retrieval ability have shifted from classical two-stage models to more comprehensive models (e.g., Rosen & Engle, 1997; Schelble et al., 2012; Unsworth et al., 2011).

A model proposed by Rosen and Engle (1997) includes four components of retrieval ability. This model assumes that despite sampling of categories and representatives, also monitoring of the retrieved representatives and suppression of previously named representatives is important. Essentially, the four-components in this model are spreading activation, monitoring, suppression, and self-generation of category cues. In contrast to more emphasis on automatic processes in previous models, these components were argued to depict not only automatic but also strategic search components (Rosen & Engle, 1997; Unsworth et al., 2011, 2013). Relatedly, the components were deemed essential for creative performance (e.g., Silvia et al., 2013). Importantly, the predictive power of the components proposed by Rosen and Engle (1997) can be tested empirically by establishing independent measures of all four components. These independent measures are best understood as cognitive ability constructs. The four proposed components (Rosen & Engle, 1997) and the cognitive abilities they might present are summarized in Table 1. Next, we explain these components in more detail and derive hypotheses regarding their implications of the nomological net of Gr.

Table 1

Four Component Model of Gr (Rosen & Engle, 1997) and Associated Cognitive Ability Constructs

Component	Explanation	Associated Cognitive Ability Construct
Automatic Spreading Activation	Automatic associative process that spreads from the cue of an item (e.g., the category "animals") to related representatives of this category and then from instance to instance (e.g., "dog" then "cat")	Crystallized Intelligence & Mental Speed
Monitoring	Monitoring of previously retrieved representatives	Working Memory Capacity
Suppression	Suppression of responses that have already been retrieved (i.e., avoiding repetitions)	Cognitive Control & Working Memory Capacity
Self-generation of cues	Finding new categories through associational processes	Secondary Memory & Crystallized Intelligence

Component 1: Automatic Spreading Activation

Retrieval of, for instance, animals or specific words is only possible when representatives are readily stored and available in long-term memory. Similarly, automatic spreading activation is only possible if there is a structure of information available where the spreading activation can take place. In addition to that, the automatic spreading activation process was argued to be associative in nature (Schelble et al., 2012). As such, a presupposition for the component of automatic spreading activation is that some sort of interconnected crystallized information structure is existent in long-term memory. In the case of verbal fluency tasks, this structure might be understood as semantic memory (Tulving, 1972), which might be conceived as a network of interconnected information (Collins & Loftus, 1975; Smith et al., 1974). Therefore, the component of automatic spreading activation should be akin to what is needed to perform well on typical Gc tasks such as declarative knowledge or vocabulary tests. Both Gr and

Gc tasks require to percolate the required information from long-term memory. In both cases, test takers are cued with an item stem (e.g., "What is the capital of Sweden?" vs. "animals"), which arguably triggers some associational processes in memory. During these processes possible responses are sampled and compared to one another in an associative manner. Taken together, performance on Gc indicators should be a meaningful predictor of Gr because prior knowledge is a necessary condition for retrieval, and tasks of both constructs tap similar underlying information structures (Jewsbury & Bowden, 2017). Gc should be equally important for short and long retrieval tasks because both tasks instantiations require the same underlying structure of information that can be retrieved from.

In addition to that, the speed of activation spreading might be associated with the processing speed of information, or mental speed (e.g., Carroll, 1993; Danthiir et al., 2005). In retrieval tasks, the relevance of speed of information processing (e.g., Danthiir et al., 2005) is apparent because test takers are required to name as many instances of a given category as possible; a requirement that is directly tied to the time constraints that is intrinsic for fluency tasks. Thus, mental speed enables fast information processing, increases the rate of retrieval per time unit (i.e., improves performance), and should therefore be deemed an important predictor of Gr (Forthmann et al., 2019; Gordon & Kindred, 2011; Jewsbury & Bowden, 2017). Furthermore, the idea of mental speed, namely carrying out supposedly simple mental processes efficiently in a given time (Danthiir et al., 2005; Goecke et al., 2021) is akin to the requirements of Gr tasks. Accordingly, recent interpretations of Gr are akin to those of mental speed, namely both being indicative of a profoundly basic level of cognitive processing efficiency (Bryan et al., 1997; McGrew et al., 2023). Due to the inherent aspect of speededness in shorter retrieval tasks, mental speed should be less important in long retrieval tasks.

Component 2: Monitoring

The monitoring of already retrieved instances (i.e., single animals) puts load on working memory (c.f., Rosen & Engle, 1997; Schelble et al., 2012). This is, because already retrieved representatives of a given category need to be kept in mind so that test takers avoid repetitions and omissions as the retrieval process progresses across the limited retrieval time. In addition, monitoring of already retrieved representatives requires an

attentional focus on the category structure (e.g., the mental representation of mammals and reptiles) while the retrieval of additional elements takes place. This can be particularly demanding, as it requires test takers to keep track of which elements have already been retrieved and possibly also which elements are still pending while at the same time making sure that they are retrieving additional elements from the correct category. Hence, monitoring might become more difficult as the time of a retrieval task progresses. Longer task times should increase the demands on WMC because, during task completion, each additional instance of a category that is generated relative to short administration times adds load on WMC. The idea of monitoring is also reflected in a typical definition of working memory, which states that working memory is used to "mentally maintain information in an active and readily-accessible state, while concurrently and selectively processing new information" (Conway et al., 2008, p. 3). In fact, the limited capacity of working memory (Baddeley, 2012; Conway et al., 2008; Cowan, 2005) has been theorized to not only account for individual differences in more complex cognitive abilities like reasoning (e.g., Kane et al., 2005; Kyllonen & Christal, 1990; Oberauer et al., 2005) but also in Gr (e.g., Rosen & Engle, 1997). As working memory is assumed to store information and control cognitive processes (Shah et al., 1999), the role of working memory in retrieval performances seems vital. Gr and WMC should be associated because WMC is essential during learning and also during retrieval.

Component 3: Suppression

Ongoing monitoring of responses is closely related to the idea that test takers need to suppress any responses that they have already retrieved and named before. This controlled suppression might be a part of the concept of "cognitive control", which is, in turn, suggested constituting a key feature of working memory (e.g., Engle, 2002). The concept of cognitive control has been reconsidered by Unsworth et al. (2011), who, among others, argue that inhibitory processes are not as crucial as previously thought (c.f., Azuma, 2004; Perret, 1974).

Component 4: Self-Generation of Cues

Self-generation of cues, or subcategories of a given cue (e.g., animals), was also suggested as an important driving mechanism for retrieval performance. The idea that the generation of new categories is based on associational processes can be linked to

secondary memory as a form of controlled strategic search (Wilhelm et al., 2013). The distinction between secondary memory and primary memory refers to dual store models of WM (Unsworth & Engle, 2007). Primary memory is understood as a unit that is responsible for the active maintenance of information for brief periods of time and the immediate recall of information. In contrast, secondary memory refers to long term storage of information and its access. It can therefore be understood as a form of controlled retrieval. Secondary memory is critical as soon as performance on primary memory, or working memory, reaches its capacity limit (e.g., Mogle et al., 2008). In terms of Gr, secondary memory might thus be relevant when test takers have exhausted one or more subcategories and need to sample representatives from a previously not used subcategory. As such, the relative importance of secondary memory should increase as time on retrieval task progresses.

Taken together, the above-described component model enables us to derive hypotheses regarding important predictors of individual differences in Gr. Overall, Gr is described as "multifaceted", and Gr performance might predominantly reflect a combination of different processes (Unsworth et al., 2011, p. 463) or, more precisely, a linear combination of established abilities.

Empirical Findings Regarding the Nomological Net of Broad Retrieval Ability

Four determinants should add to individual differences in Gr: Crystallized intelligence (Gc), mental speed, working memory capacity (WMC), and secondary memory (or long-term memory). Next, we will summarize the empirical literature on the associations of these constructs and Gr.

Crystallized Intelligence

Strong evidence suggests a vital role of Gc for Gr (Ardila et al., 1998, 2000; Jewsbury & Bowden, 2017). Latent variables for vocabulary knowledge and verbal fluency were still meaningfully related even after controlling for differential effects of age and processing speed in an age-heterogenous sample (β = .32; Hedden et al., 2005). Other evidence endorses this relation (r = .30, Hakstian & Cattell, 1978; r = .36, Unsworth et al., 2011) and established that controlling for WMC also does not diminish the role of Gc (β = .25 and β = .31, respectively; Unsworth et al., 2011; Unsworth, 2019 re-analyzing data of Hedden et al., 2005). Additionally, several other studies report and emphasize the role of vocabulary

for verbal fluency (e.g., Gordon & Kindred, 2011; Shao et al., 2014). Gc in these studies was measured by vocabulary tests but more recent concepts of Gc propose declarative fact knowledge tests as complementary Gc measures (Wilhelm & Schroeders, 2019). Given vocabulary and fact knowledge measures of Gc are very strongly related (Schipolowski et al., 2014), their relation with Gr should be next to indistinguishable. Interestingly, comparing open-ended and closed fact knowledge response formats (i.e., varying the retrieval demands in Gc tasks) did not reveal meaningful differences in correlations in a recent study (Goecke et al., 2022), which underlines the idea that Gc and Gr should be related independent of task instantiations.

Mental Speed

Mental speed and Gr showed varying strengths of relations with each other in latent variable analysis (r = .67, Forthmann et al., 2019; r = .30, Hakstian & Cattell, 1978; r = .66, Hedden et al., 2005; r = .27, Unsworth et al., 2011). A recent meta-analytic study found evidence for a medium relationship between mental speed and Gr (r = .33, corrected for attenuation; Miroshnik et al., 2023). While mental speed appears to be relevant for Gr, the meta-analysis suggests that it may not be the only factor contributing to Gr and creative idea production (Miroshnik et al., 2023). Additionally, age-associated declines in verbal fluency were found to be mainly driven by declines in the speed of information processing (Bryan et al., 1997) but not vocabulary knowledge (Gordon & Kindred, 2011; Sauzéon et al., 2011). Importantly, the predictive power of mental speed towards Gr seems to decrease once other predictors, such as WMC or Gc, are considered, but the available evidence is not conclusive (β = .15 and β = .26, respectively; Unsworth et al., 2011; Unsworth, 2019 re-analyzing data of Hedden et al., 2005). Further, relations with mental speed can also be interpreted quite differently. A reductionist perspective would claim that mental speed is indeed causal for individual differences in Gr – just like it is responsible for individual differences in other cognitive abilities (e.g., Eysenck, 1987; Jensen, 2006). From an artifact perspective, the relation between mental speed and Gr could simply be due to our inability to create task paradigms that do not share critical features, such as the time-constraints typically imposed on mental speed and Gr tasks.

Working Memory Capacity and Secondary Memory

There is ample evidence of WMC being associated with Gr, and the role of WMC for controlled search processes in verbal fluency tasks has gained considerable attention (Amunts et al., 2020; Rosen & Engle, 1997; Shao et al., 2014; Unsworth et al., 2013; Whiteside et al., 2016). For example, WMC was strongly correlated with verbal fluency scores (r = .55 and r = .43, respectively; Unsworth et al., 2009, 2011) also after controlling for differential effects of age and processing speed in an age-heterogenous sample (β = .68, Hedden et al., 2005). The predictive power of WMC towards Gr also withstood control of vocabulary and processing speed (Unsworth et al., 2011; Shao et al., 2014). However, the relationship between WMC and fluency vanished when mental speed, vocabulary, and long-term memory measures were added as predictors in another study (β = .10, Unsworth, 2019 re-analyzing data of Hedden et al., 2005). Additionally, another study concluded that WMC is only meaningfully related to verbal fluency, once other core executive functions are not partialled out of the fluency performance as the linkage between WMC and fluency would be mainly driven through their shared fundament in executive functioning (Shipstead et al., 2016, p. 788).

In contrast to the evidence of the association between working memory and Gr, there are no studies explicitly studying the relation between secondary memory and Gr jointly. However, typical tests of long-term memory (LTM; e.g., paired associates or free-recall tasks; c.f., Unsworth, 2019) are akin to measures of secondary memory. Episodic long-term memory tasks were found to be related to retrieval (Cohen, 1984; Hakstian & Cattell, 1978; Hedden et al., 2005). The relation between latent variables for free recall tasks and verbal fluency measures was estimated at r = .75 (Hedden et al., 2005). Although reduced, a significant relation between LTM and Gr remained after controlling a number of maximal cognitive effort factors (β = .23, Unsworth et al., 2019, re-analyzing data of Hedden et al., 2005).

The Present Study

The aims of the present study were twofold. First, we aimed to examine the factor structure of Gr and thereby extend previous research. Specifically, we investigated whether or not the factor structure of verbal Gr tasks (including ideational, word, and expressional fluency as narrow abilities) is affected by retrieval tasks being administered

with relatively shorter as compared to longer retrieval times. If retrieval over longer periods of time affects the nature of retrieval, this should become apparent by an overarching latent factor being insufficient to account for the covariance structure of verbal Gr tasks. If that were the case, additional nested factors that account for uniqueness due to administration times should be required. Our first research question, therefore, is: Are specific retrieval conditions (long versus short time periods) psychometrically relevant? In addition to that, we aimed to investigate how figural fluency fits into a measurement model of verbal Gr, assuming that despite the difference in content (verbal vs. figural), figural fluency can be established as another latent factor.

Second, we extended the derived measurement model for Gr with several ability factors. Specifically, we regressed Gr on Gc (including vocabulary and declarative knowledge), mental speed, WMC, and secondary memory. Based on the above-described linkage between cognitive ability constructs and cognitive components driving individual differences in Gr, we expected to find incremental predictive validities of all covariates towards Gr. To test this assumption, we captured the predictor space with a nested factor model to facilitate testing the incremental predictive validity of each covariate. Based on available evidence, we predicted to find meaningful paths between all of the specified factors and Gr.

The research objectives of the present study were not preregistered.

Method

Participants and Procedure

Data were collected in two German cities as part of a larger multivariate study which was advertised on social media, through mailing lists, via flyers, and ads in newspapers and public transportation. Participants had to be between 18 and 45 years, as cognitive decline is known to predominantly occur at later stages of life (Hartshorne & Germine, 2015). Participants were required to be native German speakers. Tasks and instructions were administered computerized on 27″ screens via Inquisit 6 (*Inquisit 6, Version 6.5.2*, 2022). Order of tasks was fixed for all participants. For all tasks, standardized instructions were presented on screen, and a trained proctor supervised each session. The computerized test battery was administered in group sessions of up to eight participants at a time. Completion of all tasks took approximately five hours including 3 breaks (2x 5

minutes and 1x 15 minutes) in which participants were provided with snacks and drinks. All participants provided written informed consent and received monetary compensation for their participation (55 Euros). We conducted the study commensurate to the ethical guidelines provided by the EU General Data Protection Regulation (*GDPR*). Anonymity of participants was guaranteed and a local ethics committee decided that the study (No. 455/21) does not require an ethic vote based on the declaration of Helsinki, as no sensible data regarding health or sexuality was gathered.

We determined the sample size for our study a priori (using *simsem;* Pornprasertmanit et al., 2021), based on considerations of several parameter estimates for our target latent factor models (i.e., minimally expected effect sizes and typical loadings of indicators). Across various competing models, we found that a sample size of $N = 300$ yields sufficient power (> .80, $\alpha = .01$) for all targeted parameter estimates (including factor loadings and factor correlations). In total, $N = 335$ participants from the general population took part in our study. Participants were excluded during data cleaning if they a) had major technical errors in data recording ($n = 2$) or b) had terminated the study early and withdrawn their consent due to personal reasons ($n = 2$). The final sample eligible for statistical analysis thus consisted of $N = 331$ participants. These participants were 65.4% female and ranged in age from 18-42 years ($M_{age} = 26.4$ years, $SD_{age} = 5.4$). $N = 263$ of the sample indicated holding at least a high school degree, and $N = 85$ indicated holding a college degree.

Measures

*Broad Retrieval Ability (Gr)***.** Gr was measured with both verbal and figural tasks. The verbal task tapped the narrow Gr abilities of ideational fluency (IF), expressional fluency (EF), and word fluency (WF). Essential features of these tasks are summarized in Table 2. All tasks were presented to participants similarly: first, participants were instructed regarding task-specific requirements. For example, instructions for IF stated that the task was to name as many representatives of a given category as possible. Then, participants were presented with an example category and example responses. Participants then had to type their responses into a textbox and press "Enter" to submit single responses until the time limit of the task was reached. For each verbal fluency task,

we administered several items (i.e., categories) in both a "long" and a "short" time condition (i.e., column "condition" in Table 2).

Table 2

Indicators of Broad Retrieval Ability (verbal).

Indicator (Narrow Gr Ability)	Condition	Item	ICC3k	*M* (*SD*)
Things_l1 (IF)	8min (long)	Animals	1	70.40 (20.84)
Things_l2 (IF)	8min (long)	Groceries	1	72.03 (24.95)
Things_s1 (IF)	1.5min (short)	Countries	.97	11.26 (4.06)
Things_s2 (IF)	1.5min (short)	Furniture	.99	12.58 (5.49)
Things_s3 (IF)	1.5min (short)	Plants	1	22.20 (7.47)
Words_l1 (WF)	8min (long)	First letter "S"	1	75.74 (21.65)
Words_l2 (WF)	8min (long)	Last letter "N"	1	71.03 (28.04)
Words_s1 (WF)	1.5min (short)	First letter "D"	1	19.53 (5.52)
Words_s2 (WF)	1.5min (short)	Last letter "R"	1	13.77 (5.18)
Words_s3 (WF)	1.5min (short)	Four letters	1	16.76 (5.69)
Synonyms_l1 (EF)	4min (long)	lonely	.82	6.08 (3.04)
Synonyms_s1 (EF)	1.5min (short)	to relax	.95	6.51 (2.65)
Synonyms_s2 (EF)	1.5min (short)	clean	.98	6.72 (3.10)

Note. All tasks were administered in German. IF = ideational fluency, WF = word fluency, EF = expressional fluency.

Figural fluency tasks (FF; e.g., *"Draw as many objects as you can based on a circle and a rectangle")* were sampled from the Berlin Structure-of-Intelligence test for Youth (Jäger et al., 2006). We employed four figural fluency tests that were assessed using paper-pencil, as these measures required participants to draw figures. Scoring for the figural fluency tasks adhered to recommendations provided in the test manual.

All retrieval tasks were rated by two or three independent human raters with respect to the number of correctly named representatives or the number of drawn figures (i.e., fluency). Agreement among raters was estimated with ICCs (Koo & Li, 2016). ICC3k

> .9 can be considered very good; and due to the pure fluency score being a sum score, a high inter-rater reliability was to be expected for all tasks. The inter-rater reliabilities for each of the verbal retrieval tasks can be found in Table 2. The inter-rater reliabilities for the four figural fluency tasks with $n = 3$ raters ranged from .98 – 1.

*Crystallized Intelligence (Gc): Vocabulary***.** Vocabulary was measured based on a selection of items from three multiple-choice vocabulary tests. We administered 50 multiple-choice vocabulary items. Twenty of these items were based on a corpus-derived vocabulary (Schroeders, 2018). Thirty items were selected from two published vocabulary tests (15 items per test; Lehrl, 2005; Schmidt & Metzler, 1992). Item compilation from these three tests was based on either empirical or expected difficulties.

35 of the 50 items consisted of five words (one real word and four non-existent words), and participants were to identify the real word. The remaining 15 items had six response alternatives. All items were scored dichotomously (1 = correct, 0 = not-correct). We computed one parcel per subscale after establishing unidimensionality for each vocabulary scale (Little et al., 2002). These parcels were used for the latent variable modeling. We excluded 3 items from the first scale (Schroeders, 2018) due to non-significant loadings, 2 items from the second scale (Lehrl, 2005) due to non-significant loadings, and were able to keep all items from the third scale (Schmidt & Metzler, 1992).

*Crystallized Intelligence (Gc): Declarative Knowledge***.** We assessed declarative knowledge using 96 items covering four broad knowledge domains and 21 subdomains, namely natural sciences (physics, biology, chemistry, mathematics, ecology), social sciences (politics, law, economy, finances), humanities (art, literature, music, religion, philosophy, history, anthropology, architecture), and life sciences (medicine, nutrition, health, domestic science) (c.f., Steger et al., 2019). Items were sampled from a large item pool (Steger et al., 2019). Half of the items were administered in multiple-choice response format with four response alternatives. The other half was administered in an open-ended response format in which participants were only presented with the knowledge question and a text box for typing in the response. Both halves were administered separately. For both tests, participants had a time limit of 18 minutes respectively. Items were presented in random order for each participant. All participants were instructed to guess the correct response in the multiple-choice format if they did not know the answer. Further, if they

did not know an answer in the open-ended format, they were instructed to indicate this by filling in "I don't know" or a question mark.

We built parcels to keep complexity of structural models presented later at a feasible level. After establishing unidimensionality for each broad knowledge domain for each response format (i.e., 8 parcels). We excluded items based on low item difficulties (i.e., < .25), low corrected item-test-correlations (i.e., r_{it} < .18), non-significant factor loadings, or considerable contribution to model misfit. In total, we excluded five items of the multiple-choice scale and six items from the open-ended scale. We then built parcels maintaining domain specificity according to the broad knowledge domains (Cole et al., 2016).

*Mental Speed***.** Mental speed was measured with three tasks. A figural comparison task (Goecke et al., 2021; Schmitz & Wilhelm, 2016) which consisted of two blocks with 40 trials each (plus 2 warm-up trials per block) used two strings of three elements (abstract figures) that were presented horizontally on the screen simultaneously. Prior to presentation of the two stimulus strings, two fixation crosses were displayed for 400 ms at positions where the stimuli would appear. Participants had to indicate whether or not strings of stimuli were exactly identical by pressing a right or a left response button respectively. Stimuli remained on the screen, until participants responded.

A numerical search task (Schmitz et al., 2016) consisted of two blocks with 40 trials each (plus 2 warm-up trials per block). Single digits ranging from 0-9 were displayed in the middle of screen successively. Participants had to decide for each stimulus whether it was a target (a "3") or not (all digits unequal "3"). They decided by pressing the left or right response key.

Lastly, an adapted version of a published lexical decision task (c.f., Lepore & Brown, 2002) used German stimuli that were based on an unpublished diploma thesis (Vurgun, 2014). Participants completed 12 warm-up trials and 120 trials. Single words were presented in the middle of the screen (250 ms), and participants decided for each word whether it is a real word or not by pressing the left or right response key accordingly. Stimuli belonged to one of four word classes: reals verbs, real nouns, fake verbs, fake nouns (Meyer & Schvaneveldt, 1971).

All mental speed tasks used a binary response format. Warm-up trials were not analyzed. The mental speed data were checked for outliers and implausible values prior to further statistical analysis. We applied the liberal Tukey criterion (Tukey, 1977), i.e., single responses were excluded when they were more extreme than 3 interquartile ranges either above the 75th or below the 25th percentile of the RT distribution, or below 200ms. We computed task scores (mean response times) that were later on used as indicators for latent variable modeling.

*Working Memory Capacity***.** WMC was measured by a location-letter binding task , a numerical updating task, and a spatial-figural updating task (Wilhelm et al., 2013).

The spatial-figural updating task used 9 differently colored rectangles that could be presented in a 3 x 3 grid matrix. Load levels for the trials varied between 2 and 5 colored rectangles. Participants completed 2 practice trials before completing 11 test trials. In a single trial, a number of colored rectangles (depending on the load level) was initially presented in the grid, followed by subsequent presentations of single colored rectangles at other positions. Participants were required to keep track of the last position of each colored rectangle. The number of position changes per colored rectangle (i.e., updating of stimuli position) ranged from 2 to 5. In the recall phase of a trial, all differently colored rectangles of a current trial were presented one at a time, and participants were asked to indicate where the target stimulus was presented last. Presentation times in the learning phase varied depending on load level: the colored rectangles were presented one at a time at a rate of either 2000ms for load level 2, 2400ms for load level 3, or 2800ms for load levels 4 and 5. Inter-stimulus-intervals lasted 1000ms. During the recall, there was no time deadline.

In the numerical updating task, numbers were presented in rectangles on the screen. Load level ranged between 2 and 6 numbers. Participants completed 2 practice trails before completing 12 test trials. Within one trial, a number of rectangles (depending on the load level) was initially presented, together with corresponding numbers (for 3000 ms). After that, the numbers disappeared from the rectangles and were then updated rectangle by rectangle (pseudo-randomly). Participants were required to keep track of the last presented number per rectangle and to type the number for each rectangle at the end of each trial. The amount of numbers updated per rectangle ranged from 2 to 6. The

numbers were presented one at a time at a rate of 1600ms for all load levels. Inter-stimulus-intervals lasted 1000ms. During the recall, there was no time deadline.

The location-letter binding task used letters at different positions on a 3 x 3 grid as stimuli. A trial consisted of the sequential presentation of a list of letter-position bindings, which participants were asked to memorize. Instantaneously after the presentation phase, a recall test of each letter-position binding followed in random order. For 50% of the recall trials, a position in the grid was marked, and participants were required to select the letter that previously had been paired with the presented position. For the other 50% of the recall trials, the presentation was converse. Participants were presented with one of the letters and had to choose the position in the grid where the letter was presented before. Load levels of this task ranged from 2 to 6 pairs. Participants completed 2 practice trials before completing 14 test trials. During the learning phase, the presentation time for each letter-position binding was held constant across trials and amounted to 1500ms.

Scoring of the working memory capacity tasks applied a partial credit scheme (Conway et al., 2005). For all tasks, accuracy was computed as the relative proportion of correct responses across queries in a trial. Subsequently, the computed mean across trials formed the task scores that were used as indicators in the latent variable analysis later on.

*Secondary Memory***.** Secondary memory (SM) was measured with a word-word and a word-number task (c.f., Wilhelm et al., 2013). In the word-word task, participants learned 20 word pairs sequentially. In the word-number task, participants memorized 20 word-number (two-digit numbers) pairs sequentially. Presentation time for each stimulus pair was 4000 ms, followed by an interstimulus-interval of 1000 ms. There were two blocks of trials, each consisting of 20 pairs, for both tasks. The two blocks per task included different lists of stimuli, and they were located at different positions within the test battery. Between encoding and recall, participants completed other tasks for 3 minutes to ensure they could not retain some of the pairs in WMC. After the retention interval, participants were asked to recall the learned stimulus pairs of each list, again, at different positions within the test battery. During the recall phase, there was no time limit. For both tasks, participants were either provided with the first or second stimulus of a respective pair and then had to type in the missing associate with the keyboard.

Scoring of secondary memory tasks also applied a partial credit scheme (Conway et al., 2005). For all tasks, accuracy was computed as the relative proportion of correct responses.

*Typing Speed***.** We measured typing speed because it can be deemed a nuisance variable in the computerized assessment of verbal retrieval abilities (Forthmann et al., 2017). Participants were asked to type sentences shown on screen as fast as they could. All participants typed four sentences (see SM Page 1). Participants saw a fixation cross followed by a sentence in the middle of the screen, together with a textbox underneath it. Participants could immediately start to type a sentence. One sentence was a pangram (i.e., a sentence using all 29 German letters). The other three sentences were shorter and included typical German words. Typing times for correct responses (i.e., matching exactly the provided sentences after decapitalizing all characters) were recorded in milliseconds. Accuracy in this task was at 80%. Erroneous responses were set to missing.

Data Treatment

We checked for multivariate outliers across tasks based on the Mahalanobis distance (Meade & Craig, 2012). The Mahalanobis distance is the standardized distance of one data point from the mean of the multivariate distribution of a multivariate set of manifest variables. For our study, we considered participants' fluency scores of three of the retrieval indicators (things long, words long, and synonyms long) and the scores of the three vocabulary tests to account for the verbal nature of most of the administered tasks. We excluded $N = 9$ participants due to their Mahalanobis distances > 20. After data cleaning, our final sample thus included $N = 322$ participants.

Statistical Analysis

All analyses were conducted with R (R Core Team, 2022) and we provide all materials necessary to reproduce the analyses in an online repository: https://osf.io/erdxq/?view_only=adec718e1ab64803be9892e5c9b0ff06

Confirmatory Factor Analysis (CFA) was carried out by the R package *lavaan* (Rosseel, 2012). Where possible, we used *full information maximum likelihood* estimation under the assumption of missing completely at random to combine missing data and parameter estimation in a single step (Enders, 2010; Schafer & Graham, 2002). Less than 0.4% of the data were missing. Models based on dichotomous indicators were estimated

based on the *Weighted Least Squares Means and Variance* adjusted (WLSMV) estimator (Beauducel & Herzberg, 2006); in this case, pairwise observations were used. Models based on continuous indicators are based on maximum likelihood estimator with robust standard errors (MLR). The following fit statistics were considered indicating good model fit: CFI (*Comparative Fit Index*) ≥ .95, RMSEA (*Root Mean Square Error of Approximation*) ≤ .06, and SRMR (*Standardized Root Mean Square Residual*) ≤ .08 (Hu & Bentler, 1999). For acceptable model fit these boundaries were used: CFI ≥ .90, RMSEA ≤ .08, and SRMR ≤ .10 (Bentler, 1990; Browne & Cudeck, 1992). We used McDonald's ω as indicator of factor saturation (McDonald, 1999; Raykov & Marcoulides, 2011). Factor saturation indicates how much variance is accounted for by a latent variable in all underlying indicators (Brunner et al., 2012). Factor saturation was computed according to Brunner and colleagues (2012).

Prior to modeling, indicator variables were standardized in order to account for differences in scales. Correlations, means, and standard deviations of all variables is provided in the supplementary materials (SM Table 1).

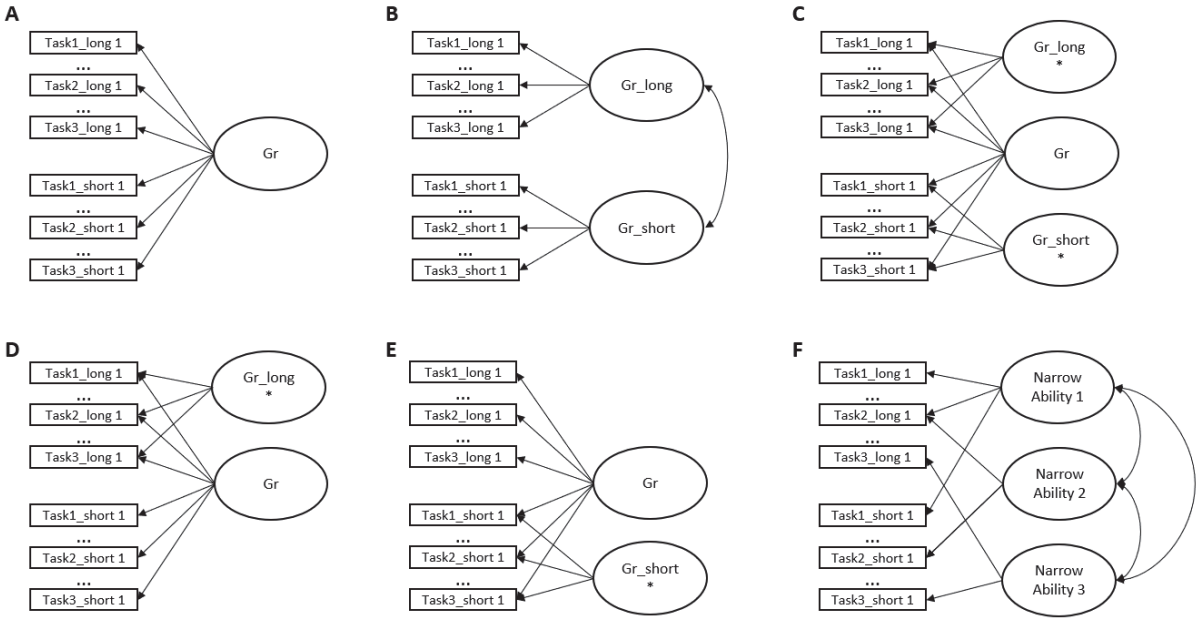
Results

Are We Able to Distinguish Between Short and Long Periods of Retrieval in the Verbal Domain?

We were interested in whether short and long periods of retrieval could be distinguished psychometrically in the administered verbal fluency tasks tapping ideational, word, and expressional fluency (i.e., narrow abilities). Therefore, we compared competing measurement models addressing this research question (see Figure 1 for a schematic overview of the theoretical models).

Figure 1

Conceptual Measurement Models of Verbal Gr Accounting For Short and Long Retrieval Time Periods.



Note. Schematic depiction of competing measurement models of Gr. * indicate a different interpretation of the according latent factor as compared with the other models.

Model A proposes a general factor reflecting Gr and therefore corresponds to the idea that different time periods and different tasks for retrieval play no role in individual differences in Gr. Model B postulates two correlated factors, which are indicated by long and short retrieval tasks, respectively. Therefore, the model assumes that long and short periods of retrieval indicate two distinct but correlated psychometric constructs. In model C, a full bifactor model is proposed where a general factor of Gr is specified, along with two factors capturing the specific communalities of long and short tasks, respectively, nested underneath the general factor. Model D and model E are variations of model C (c.f., bifactor-S-1 models, e.g., Eid et al., 2017, 2018), where either only the specificity above and beyond a general factor of Gr is modeled in long retrieval tasks (model D) or in short retrieval tasks (model E). As such, these models assume that only one time condition is subject to individual differences not accounted for by the general factor. Lastly, model F specifies three task-specific correlated group factors, which refers to the assumption that the narrow abilities which can be assigned to the tasks (ideational fluency (things), word

fluency (words), and expressional fluency (synonyms)) can be distinguished psychometrically. The fit indices of all tested measurement models can be found in Table 3.

Table 3

Fit Indices of Measurement Models of Verbal Gr Accounting For Short and Long Retrieval Time Periods.

	Measurement Model	χ^2	df	CFI	RMSEA	[90% CI]	SRMR	ω
A	g-factor	387.27	65	.845	.124	[.113; .136]	.066	Gr = .91
B	Correlated factors long and short	384.49	64	.846	.125	[.113; .137]	.066	Gr_long = .87; Gr_short = .84
C	Full Bifactor long and short nested	291.21	52	.885	.120	[.106; .134]	.055	Gr = .97; Gr_long = .24; Gr_short = .00
D	Bifactor S-1 long nested	371.38	60	.850	.127	[.115; .140]	.066	Gr = .97; Gr_long = .00
E	Bifactor S-1 short nested	321.59	57	.873	.120	[.107; .133]	.055	Gr = .86; Gr_short = .00
F	Correlated factors narrow abilities	154.6	62	.956	.068	[.055; .081]	.040	IF = .87; WF = .85; EF = .77

Note. All models were based on $n = 13$ indicators and on a sample of $n = 322$ participants. The correlation between Gr_short and Gr_long in model B was $r = .98$ and thus close to unity; Model C exhibited a negative variance in one indicator, which is a common problem in full bifactorial latent models (e.g., Eid et al., 2017; Eid et al., 2018).

Models A to E all had unacceptable fit. The measurement model best fitting to the current data is a correlated factors models, where individual differences in the narrow abilities of Gr (i.e., ideational, expressional, and word fluency) are depicted by correlated group factors (c.f., model F in Figure 1). The model negates the existence of individual differences due to different retrieval time periods. In this model, the narrow abilities of Gr (c.f., Schneider & McGrew, 2018) are specified as distinct but correlated entities. The correlations between the latent factors of ideational fluency, word fluency, and expressional fluency amounted to r = .81 (SE = .026; p < .001) between ideational fluency and word fluency, to r = .71 (SE = .040; p < .001) between ideational fluency and expressional fluency, and to r = .66 (SE = .047; p < .001) between word fluency and expressional fluency, respectively. All factors captured substantial shares of variance (all p < .001). In addition to that, all three factors showed good reliabilities (c.f., Table 3). The loadings of the single factors ranged from λ = .65 - .87 for ideational fluency, from λ = .64 - .90 for word fluency, and λ = .63 - .78 for expressional fluency (all p < .001). Detailed information regarding the loadings of the single indicators in this model are provided in SM Table 2 in the supplementary materials.

This model was extended with a latent typing speed variable as a predictor of all three correlated factors. A latent variable model of typing speed with typing times for the four sentences as indicators had sufficient fit (n = 322; $\chi^2(2) = 8.36$, CFI = .935, RMSEA = .101, SRMR = .027; ω = .87). The correlated factors model including typing speed fit the data well (χ2(113) = 245.88, CFI = .943, RMSEA = .060, SRMR = .048; c.f., SM Figure 1). We found that, as expected, typing speed was a moderate predictor of ideational, word, and expressional fluency (β_{IF} = -.42 (SE = .06), β_{WF} = -.45 (SE = .05), β_{EF} = .31 (SE = .06)), but that the correlations between the three factors were by and large unaffected by this ($r_{IF\sim WF}$ = .77; $r_{IF\sim EF}$ = .67; $r_{WF\sim EF}$ = .62). Hence, the observed correlations between the three factors cannot be accounted to individual differences in typing speed.

Please note that model F is statistically equivalent to a higher-order factor model, with a higher-order factor Gr at the apex of the hierarchy of three first-order factors. A higher-order factor model explicitly represents the overarching ability (Gr), as proposed in contemporary models of intelligence structure (e.g., Carroll, 1993; Schneider & McGrew, 2018). This model implies that individual differences in narrow retrieval abilities are

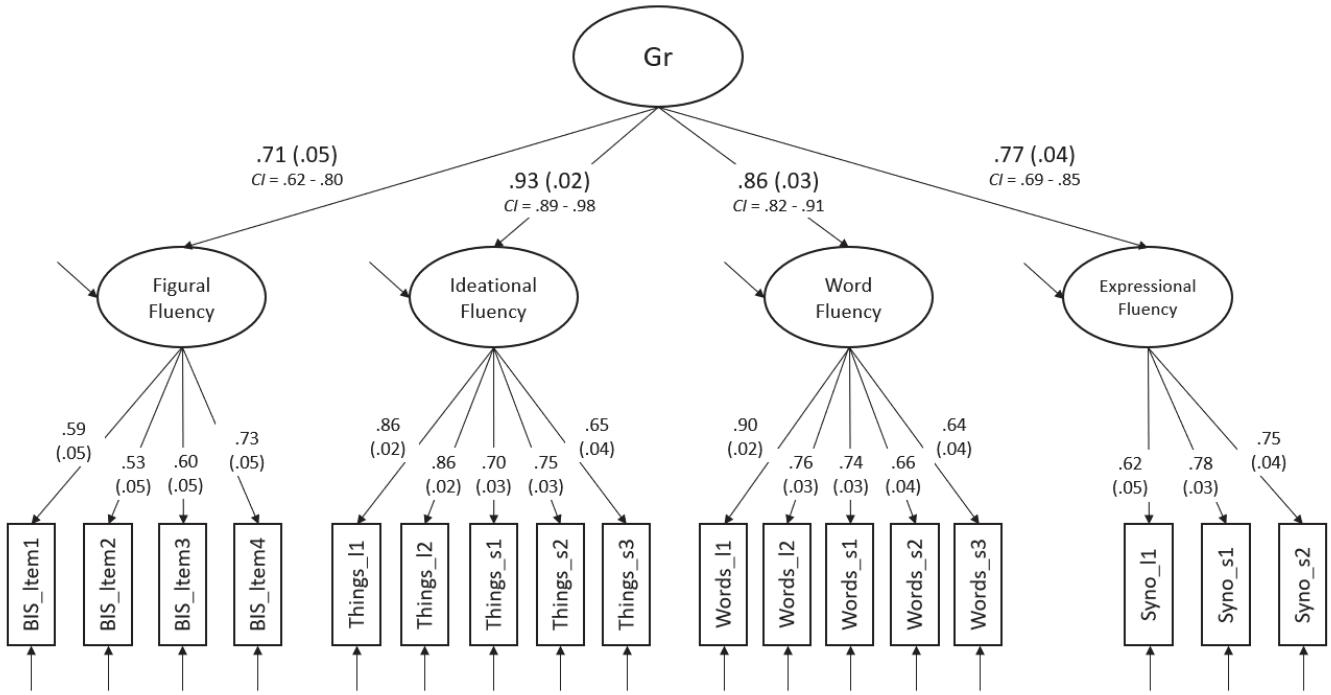
independent of time periods of retrieval. In fact, modeling Gr that specific way is in line with previous literature (e.g., Silvia et al., 2013).

Measurement Model of Gr Including Figural Fluency

After establishing a measurement model of Gr based on the narrow abilities more verbal in nature (i.e., ideational, word, and expressional fluency), we were interested in broadening the coverage of the model by adding a factor of figural fluency to it. As described above, model F (see Figure 1) is statistically equivalent to a higher-order factor model. Therefore, it made sense to add another first-order factor to the already established higher-order factor model of verbal Gr. A higher-order model with four first-order factors (see Figure 2) is more parsimonious than a correlated factors model with four factors. In this model, the general factor captures the covariance of the four latent trait factors (figural fluency, ideational fluency, word fluency, and expressional fluency). The higher-order factor model fitted the data well (c.f., Figure 2), and the higher-order factor of Gr captured substantial shares of variance ($p < .001$). The factor saturation of the higher-order factor was high ($\omega_{Gr} = .85$). The loadings of the first-order factors on the higher-order factor and the single loadings of the indicators are depicted in Figure 2. Again, the higher-order factor model fitted the data better than a unidimensional model with a single general factor supposed to account for all communality of all indicators (($\Delta\chi 2(4, N = 322) = 234.26$, $p < .001$).

Figure 2

Higher-Order Measurement Model of Gr.



Note. Syno = Synonyms, CI = 95% Confidence Interval. All parameters are standardized. Standard errors are depicted in parentheses. $n = 322$; $\chi^2(115) = 195.51$, CFI = .967, RMSEA = .047, SRMR = .037. $\omega_{Figural} = .71$; $\omega_{Ideational} = .87$; $\omega_{Word} = .85$; $\omega_{Expressional} = .77$.

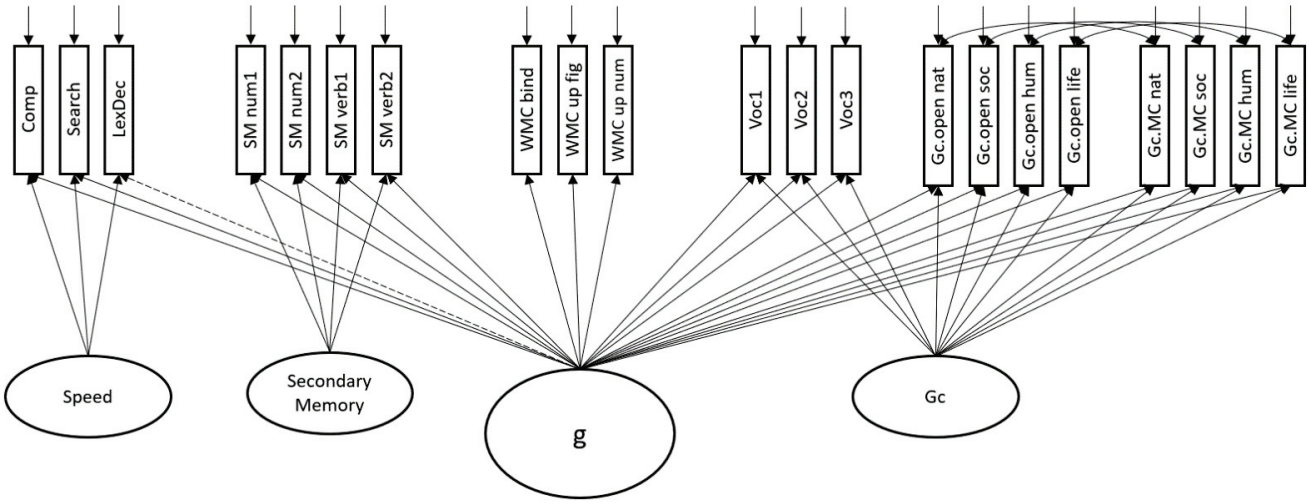
Predictors of Individual Differences in Gr

In order to test the previously derived hypotheses regarding the underlying cognitive mechanisms of Gr performance (c.f., Table 1), we aimed at regressing the higher-factor of Gr onto the covariates. Therefore, we first established a measurement model of the covariates. To this end, we specified a bifactorial measurement model of the covariates based on the indicators of working memory capacity, secondary memory, Gc (including vocabulary and declarative knowledge), and mental speed. The model included a general factor (*g;* which was marked by the WMC indicators), and three nested factors depicting secondary memory, mental speed, and Gc. The model is depicted in Figure 3. Whereas *g* accounts for the communality of all indicators (i.e., variance that is common to all indicators), the remaining specific variances in the indicators were captured by the nested factors. For example, the factor labeled mental speed depicts variance that is unique to mental speed once any influence of *g* is controlled for. In other words, the nested factors depict the variance in the indicators that is left after controlling for *g*. Due to this modeling

rationale, the model allowed for testing the specific increments of the single factors above and beyond their shared variances depicted by *g*. Therefore, this modeling approach is helpful in investigating the specific contributions of the single covariates to Gr. Taken together, the model fitted the data well (n = 322; $\chi^2(167) = 375.22$, CFI = .922, RMSEA = .062, SRMR = .064) and all factors showed significant variance (all $p < .001$) and satisfying factor saturations (ω_g = .78; $\omega_{Secondary\ Memory}$ = .47; ω_{Speed} = .59; ω_{Gc} = .50). The loadings of the single indicators are depicted in Table 4.

Figure 3

Measurement Model of Cognitive Covariates.



Note. WMC bind = Working Memory Capacity binding, WMC up = Working Memory Capacity updating, SM = Secondary Memory, Voc = Vocabulary, Comp = Comparison, LexDec = Lexical Decision, fig = figural, num = numerical, verb = verbal, nat = natural sciences, soc = social sciences, hum = humanities, life = life sciences. All loadings are standardized. The residuals of the declarative knowledge indicators were allowed to correlate freely for the same content domains in order to account for their domain specificity ($r_{nat\sim\sim nat}$ = .34; $r_{soc\sim\sim soc}$ = .27; $r_{hum\sim\sim hum}$ = .18; $r_{life\sim\sim life}$ = .36; all $p < .01$).

Table 4

Loadings of Indicators of Cognitive Covariates Measurement Model (c.f., Figure 3).

	λ (*SE*)			
Indicator	***g***	**Mental Speed**	**Secondary Memory**	**Gc**
WMC bind	.70 (.05)	-	-	-
WMC up fig	.65 (.06)	-	-	-
WMC up num	.77 (.05)	-	-	-
Comp	.20 (.07)	.52 (.07)	-	-
Search	.20 (.08)	.65 (.08)	-	-
LexDec	*.03 (.07)*	.54 (.07)	-	-
SM num1	.52 (.06)	-	.47 (.06)	-
SM num2	.53 (.06)	-	.50 (.06)	-
SM verb1	.56 (.04)	-	.75 (.04)	-
SM verb2	.58 (.05)	-	.66 (.05)	-
Voc1	.46 (.06)	-	-	.57 (.06)
Voc2	.35 (.07)	-	-	.58 (.06)
Voc3	.49 (.06)	-	-	.64 (.05)
Gc.open nat	.56 (.06)	-	-	.34 (.08)
Gc.open soc	.26 (.07)	-	-	.49 (.06)
Gc.open hum	.43 (.07)	-	-	.40 (.08)
Gc.open life	.41 (.06)	-	-	.50 (.06)
Gc.MC nat	.43 (.07)	-	-	.39 (.07)
Gc.MC soc	.34 (.07)	-	-	.23 (.07)
Gc.MC hum	.32 (.06)	-	-	.55 (.05)
Gc.MC life	.30 (.06)	-	-	.50 (.05)

Note. WMC bind = Working Memory Capacity binding, WMC updt = Working Memory Capacity updating, SM = Secondary Memory, Voc = Vocabulary, Comp = Comparison, LexDec = Lexical Decision, fig = figural, num = numerical, verb = verbal, nat = natural sciences, soc = social sciences, hum = humanities, life = life sciences. All loadings are standardized. Standard errors are depicted in parentheses. All $p < .001$, except for values printed in italics (c.f., LexDec). Please note that for ease of interpretation the loadings of the mental speed indicators were multiplied by -1 so that positive loadings reflect better performance (i.e., faster reaction times are better).

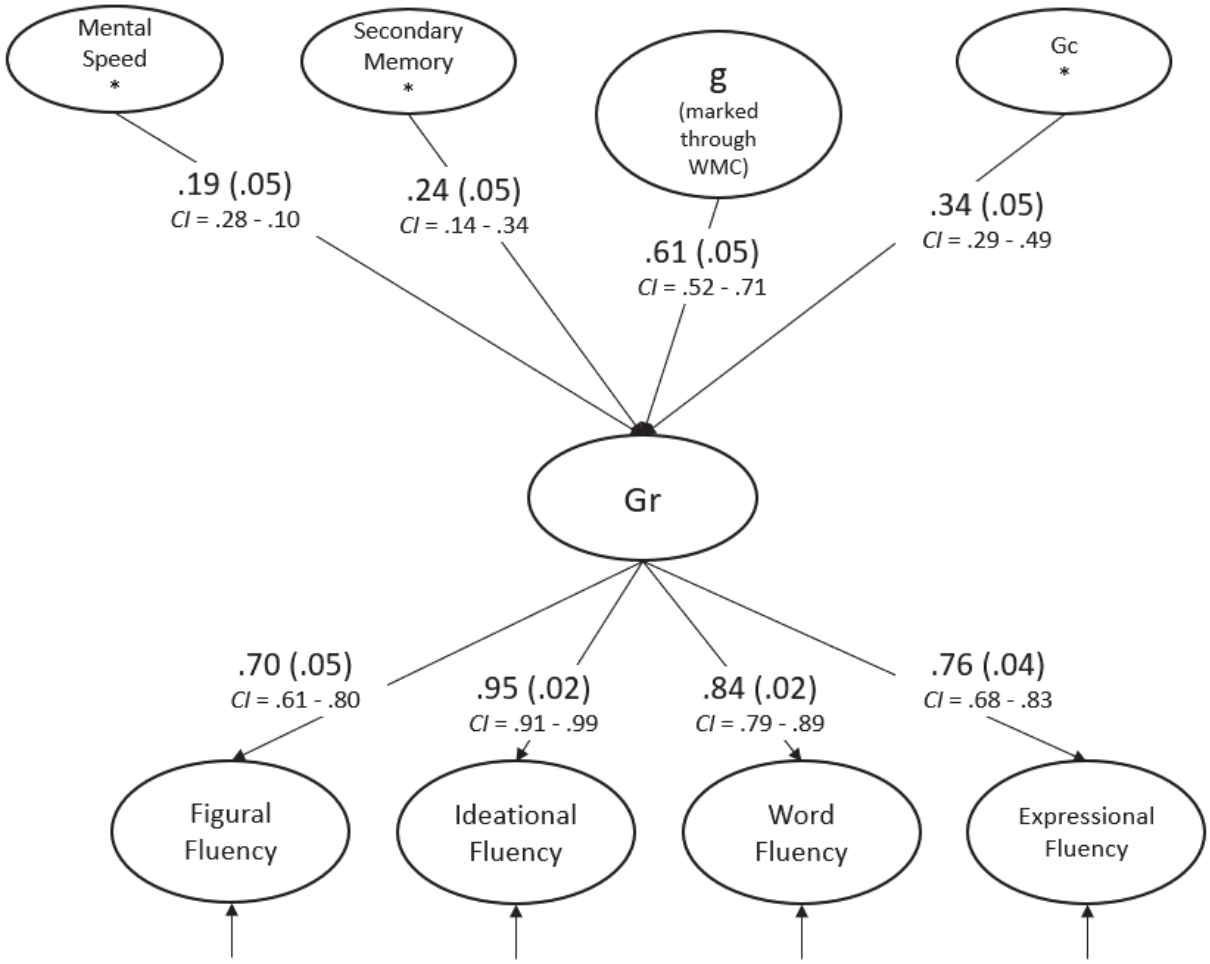
Subsequently, we combined the model presented in Figure 3 with the measurement model of Gr (c.f., Figure 2) to a latent regression model in which Gr is predicted by *g*, and the nested factors. In Figure 4, the latent regression model is depicted, including the standardized latent regression weights. We found that individual differences in Gr were strongly predicted by the *g*. In addition to that, however, also the nested factors—which capture specific variance not already accounted for by *g*—meaningfully predicted Gr

above and beyond *g*. Whereas Gc accounted for most variance in Gr after *g*, secondary memory was closely followed by mental speed. In total, 62% of variance (R^2 = .62) in Gr were explained by the four predictors.

Again, this model was extended with a latent typing speed variable, which was allowed to correlate with *g* and predicted Gr (χ2(787) = 1278.41, CFI = .920, RMSEA = .044, SRMR = .057; c.f., SM Figure 2). We found that, as expected, typing speed was still a predictor of Gr (β = -.25 (*SE* = .05)), and that typing speed correlated moderately with *g* (*r* = -.42 (*SE* = .07)). However, the height of the regression weights towards Gr were by and large unaffected ($\beta_{Gr\sim g}$ = -.51 (*SE* = .06); $\beta_{Gr\sim Mental\ Speed}$ = .15 (*SE* = .05); $\beta_{Gr\sim Secondary\ Memory}$ = .25 (*SE* = .05); $\beta_{Gr\sim Gc}$ = .38 (*SE* = .05)). In fact, controlling for typing speed increased the predictive power of Gc towards Gr. Hence, the typing speed may have contributed to individual differences in Gr, but this effect was negligible for the prediction of Gr through the covariates.

Figure 4

Latent Regression Model with Cognitive Covariates Predicting a Higher-Order Factor of Gr.



Note. A * indicates that certain factors are nested underneath *g* (c.f., Figure 3). CI = 95% Confidence Interval. All parameters are standardized. Standard errors are depicted in parentheses. $n = 322$; $\chi^2(635) = 1047.14$, CFI = .926, RMSEA = .045, SRMR = .055. Please note that for ease of interpretation the regression weight of the mental speed towards Gr was multiplied by -1 so that positive loadings reflect better performance (i.e., faster reaction times are better).

Discussion

We investigated the dimensionality of Gr by testing competing measurement models that included verbal indicators of Gr with short and long retrieval time periods and figural fluency tasks, respectively. Thus, we went beyond the scope of previous studies that mainly focused on verbal retrieval. In addition to that, we derived hypotheses regarding explanatory cognitive variables of individual differences in Gr and put these to critical tests by means of structural equation modeling. With that, we were able to test

whether Gr should be understood as more than the sum of its' parts. In the next sections, we will shortly summarize and discuss our findings and address their implications for further research.

The Measurement Model of Gr

First, we have tested a number of competing measurement models of verbal retrieval. Our results clearly illustrate that a psychometric distinction between short and long verbal retrieval tasks is not empirically tenable. Importantly, we provided evidence that strong individual differences in these factors still remain after controlling for typing speed, which could have confounded our results (Forthmann et al., 2017). Although the indicators for short and long retrieval time periods differed with regard to their item content, our latent variable approach allows abstracting from item specificities because we applied a large number of single retrieval items in all domains.

In addition to that, we extended the measurement model of verbal Gr with four indicators of figural fluency and found that a higher-order factor structure described our data best. Finding evidence for the four first-order factors of figural fluency, expressional fluency, word fluency, and ideational fluency stands in contrast to earlier research (Jewsbury & Bowden, 2017) but is in line with assumptions articulated in contemporary intelligence structure models like the CHC taxonomy (Schneider & McGrew, 2018). Moreover, our results are in line with other evidence proposing Gr as a higher-order factor (Silvia et al., 2013). Although our reported model was based mostly on verbal retrieval indicators, we provide evidence for the figural stimulus modality being a meaningful complement regarding the construct coverage of Gr. We would expect that further narrow Gr abilities (e.g., associational fluency) should be insertable into this model structure, but this remains an empirical question.

Tasks assessing retrieval ability are frequently used in applied settings like neuropsychological assessments, but only few studies considered the dimensionality of Gr. It is of practical relevance to know whether time restrictions imposed on retrieval tasks alter the construct that is measured. Our results imply that measuring Gr based on different time periods for retrieval in the single retrieval tasks is meaningless. Thus, pertaining to short versions of retrieval tasks is possible without any loss of information regarding the rank-order of individuals. This has considerable implications for applied

settings. These include fundamental savings regarding both temporal and monetary costs but also possible gains in measurement precision. As short retrieval tasks require much less administration time (in our case, short tasks took approximately 1/5 of the time of long retrieval tasks), both the costs for administration and the number of responses per respondent is decreased. In turn, this results in less cost-intensive scoring labor (e.g., Forthmann et al., 2022). Capitalizing on that, administering shorter retrieval tasks offers the possibility to measure Gr more broadly, as more single items can be assessed in a still shorter amount of time as compared to administering longer tasks. Hence, for applied settings, we recommend to rather use several shorter retrieval tasks from different first-order factors of Gr than using one longer retrieval task pertaining to one specific first-order factor of Gr.

The Nomological Net of Gr

We were interested in the individual contributions of several cognitive ability variables to a higher-order factor of Gr because, in terms of scientific parsimony, it might be argued that not much of Gr is left after accounting for these antecedents. Importantly, we linked the cognitive ability variables with components of an explanatory model of individual differences in retrieval ability (e.g., Rosen & Engle, 1997). Thus, we regressed Gr on Gc (including vocabulary and declarative knowledge), mental speed, WMC, and secondary memory in a structural equation modeling approach. Importantly, we captured the predictor space with a nested factor model to facilitate testing the incremental predictive validity of each covariate. We did so because we were interested in the specific contributions of the covariates toward Gr. The measurement model of our cognitive covariates was a bifactor-(S-1) model (Eid et al., 2017, Eid et al., 2018), as it contains three specific factors (Gc, mental speed, and secondary memory), while omitting a fourth specific factor, WMC, that served as reference for the general factor. Marking the general factor with indicators of WMC is reasonable because WMC is highly correlated with fluid intelligence (Kane et al., 2005; Oberauer et al., 2005), and fluid intelligence is the single best indicator of *g* (c.f., Wilhelm, 2005). Next, we will discuss the results of our latent regression approach sorted by the empirical relevance of the predictors for Gr.

Gr and g or WMC

We found that the general factor of the covariates (marked by WMC) explained the largest amount of variance in the higher-order factor of Gr. Although our modeling approach makes it difficult to separate the specific contribution of WMC towards Gr, we interpret this result as relevance of WMC for Gr. With that, our results replicate earlier research emphasizing the role of WMC for Gr performance (e.g., Hedden et al., 2005; Rosen & Engle, 1997; Unsworth et al., 2009, 2011).

We mapped the component of monitoring onto WMC because keeping track of retrieved representatives requires a cognitive construct that quickly encodes, stores, and updates units of information in a readily available manner. As the time to retrieve representatives from LTM increases, more monitoring needs to be performed to prevent errors. Hence, and in accordance with previous literature (e.g., Rosen & Engle, 1997; Unsworth et al., 2013), we hypothesized that the relation between WMC and Gr should increase with increased retrieval time periods because the latter arguably put more load on WMC, as more instances need to be monitored (i.e., stored and held readily available in WMC). Although we did not find evidence for this hypothesis, as short and long periods of retrieval were not to be distinguished psychometrically, we still found that WMC plays a fundamental role in individual differences in Gr.

Gr and Gc

The second largest amount of variance in Gr was explained by the nested factor of Gc. We went beyond previous research regarding Gc and Gr, as our Gc factor was indicated through a broad range of vocabulary and declarative knowledge tests in different response formats (i.e., multiple-choice vs. open-ended). Further, the reported measurement model of Gc replicates earlier research showing that tests of vocabulary and declarative knowledge share a common core (Schipolowski et al., 2014) and that response format plays no role in individual differences in declarative knowledge (Goecke et al., 2014). The predictive power of Gc towards Gr implies that Gc explains incremental variance in Gr that is not accounted for by the other applied covariates. This makes sense, as retrieval of information from long-term memory will always require some form of prior knowledge (i.e., information that is stored and readily available in LTM).

We argued that the component of automatic spreading activation can be understood as Gc, because it requires a net of available knowledge, which is akin to the concept of Gc. In fact, the information that must be retrieved from LTM in prototypical fluency tasks might be understood as a product of information that are cumulatively acquired across various episodes of life. These information, in turn, depict Gc (e.g., Wilhelm & Kyllonen, 2021). Whereas Gc should be understood as the product of fluid abilities (Cattell, 1987), an ongoing acculturation process (Ackerman, 1996) and of intellectual investment traits (e.g., von Stumm, 2013; Ziegler et al., 2012) such as openness, retrieval ability, on the other hand, might be considered the mechanic capability to access and recall all this accumulated information quickly and efficiently.

Gc requires retrieval by the nature of measurement (i.e., finding the correct response to a given question). Prototypical Gr tasks require Gc by the nature of information needed (i.e., semantic information). Whereas the intention of Gc measures is for test takers to fail due to lack of knowledge but not lack of retrieval, the sentiment is the other way around for retrieval tasks. Importantly, both constructs can hardly be conceptualized independently from one another: both rely on a priori stored information, and all knowledge is meaningless if no mechanism is in place to retrieve it. Our results endorse this interpretation because Gc was incrementally predictive towards Gr, even after the communality with other covariates was accounted for.

Gr and Secondary Memory

Third in line regarding the predictive power towards Gr was the nested factor of secondary memory. We found that secondary memory was incrementally predictive of Gr, above and beyond *g*, and the other applied covariates. As our secondary memory tasks required the cued free recall of previously learned information (i.e., word-word pairs or word-number pairs), this endorses previous interpretations of strategic search components being important for Gr performance (e.g., Rosen & Engle, 1997; Troyer et al., 1997; Unsworth et al., 2011). Importantly, this might also reflect the relative importance of learning efficiency (*Gl*; Schneider & McGrew, 2018) for Gr. Considering that only the information that were previously learned can be retrieved in the recall phase of the tasks, all secondary memory tasks reflect some sort of learning performance.

We mapped the component of self-generation of cues onto secondary memory, a memory system that arguably goes beyond ordinary working memory (e.g., Mogle et al., 2008) and is more important for goal-oriented search in LTM (e.g., Wilhelm et al., 2013). We hypothesized a relevance of secondary memory for individual differences in Gr because when test takers exhaust one subcategory, they need to explicitly search for a new subcategory that has not been activated before. Although the variance decomposition by means of our structural equation modeling approach might not allow us to speak regarding the use of new subcategories, our evidence distinctly emphasizes the role that secondary memory retrieval (i.e., goal-oriented retrieval from previously learned information) plays for less goal-oriented retrieval from LTM (i.e., typical Gr tasks).

Gr and Mental Speed

Lastly, our results suggest that mental speed explained incremental amounts of variance in Gr over and above the other covariates. We measured mental speed with three different tasks in three different stimulus domains. Thus, our mental speed factor can be deemed fairly broad. Our results endorse earlier suggestions of mental speed playing a role in Gr performance (e.g., Miroshnik et al., 2023), although the relative importance as compared to other driving factors like WMC or Gc seems less crucial.

We argued that mental speed might play a role in the automatic spreading activation component. The faster the speed of information processing, the better test takers should perform in Gr tasks because the number of retrieved elements per time unit might be increased (e.g., Forthmann et al., 2019). One argument regarding previous evidence of the relationship between mental speed and Gr pertains to the immanent task features that are shared between typical Gr and mental speed tasks; that is, their implicit speededness. This means that the correlation between mental speed and Gr, as reported in the literature hitherto, could be interpreted as an artifact of task impurity rather than an actual association between the two constructs. One might even argue that the relevance of mental speed for Gr performance should decrease with decreased speededness of the tasks because the more time is given for retrieval, the less important it becomes to retrieve all relevant information as fast as possible. However, we have shown that even if the speededness of Gr tasks is drastically decreased (i.e., long retrieval tasks with 8 minutes of time for retrieval), the relationship with mental speed remains because the variance of

short and long retrieval tasks can amazingly enough not be decomposed. In this vein, future research should investigate whether the time constraints we have imposed on the retrieval tasks at hand and classified as *short* and *long* were short and long enough. It remains an open question whether a psychometric distinction could have been found with even more contrasting timing conditions.

Is Gr More Than the Sum of Its Components?

Summing up, our results indicate that 62% of variance in Gr could be explained by the applied covariates in our study. Although this is a large amount of variance, 38% of individual differences in Gr are unaccounted for. This endorses a perspective on Gr as a strong and discriminant factor of cognitive abilities in current intelligence structure models (e.g., Schneider & McGrew, 2018) and shows that Gr is more than a linear combination of its parts. Nevertheless, we also found strong evidence for earlier explanatory attempts to disentangle individual differences in Gr.

Desiderata for Future Research

We showed that Gr is a strong and discriminant factor of cognitive abilities that is not affected by time constraints imposed on prototypical retrieval indicators. In the following sections, we want to point towards future directions and research questions regarding Gr that remain open. In addition to that, future research should replicate the present results in an effort to contribute to replicable and cumulative research (Flake & Fried, 2020).

First, it is unclear whether the remaining variance in Gr depicts real individual differences or is just a mere artifact of the circumstance that our predictor space was not exhaustive. For example, it is possible that more specifically tailored tests of Gc tapping not only broad knowledge domains (e.g., Steger et al., 2019) but more specific knowledge domains (e.g., Rusche & Ziegler, 2023), including the topics of the administered retrieval tasks (e.g., animals, furniture, etc.) would have led to larger amounts of variance in Gr being explained. Future studies should therefore consider overlapping domains regarding the measurement of Gr and Gc. As arguably both constructs require the same—or at least a highly similar—structure of underlying information, this might further inform us about their relationship.

Second, to broaden our understanding of the cognitive processes underlying recall from information from LTM, it seems important to consider alternative retrieval tasks that do not solely rely on information stored in semantic LTM. Given that the information we accumulate over the course of our lifetimes is also deeply personal (c.f., Thomsen, 2015), it might be worth considering retrieval tasks that probe episodic or even autobiographical information. This would require prudent test designs that capitalize on verifiable autobiographical information. For example, it would be interesting to study the extent to which performance on prototypical semantic retrieval tasks (e.g., naming animals) is associated with performance on retrieving autobiographical information (e.g., "Name all your classmates who attended first grade with you."). This might inform us about the generalizability of retrieval processes.

Third, we acknowledge that Gr performance does not necessarily have to be understood in terms of pure fluency scores. Although the instructions of prototypical retrieval tasks are clearly stated and usually demand test takers to name *as many* instances of a given category as possible, other performance indicators of Gr are plausible. These include, for example, scores of clustering and switching (e.g., Unsworth et al., 2011), which depict the number of categories test takers sampled from, and the number of category switches test takers performed, respectively. Although obtaining scores regarding clustering and switching is not necessarily in line with what is expected of test takers (e.g., why should the performance of a test taker who sampled from more than one category be classified better than the performance of a test taker who sampled the same number of instances from only one category?), such scores might provide important insights into the cognitive underpinnings of retrieval performances. For example, category switches could be studied in relation to actual task-set switching (e.g., Kiesel et al., 2010; Rogers & Monsell, 1995). In fact, experimental designs manipulating category switches in retrieval tasks (e.g., in an AABB-design, c.f., Kiesel et al., 2010) and thereby increasing the executive demands of retrieval tasks could extend our understanding of the role of variables associated with terms like cognitive control for retrieval performance.

Relatedly, inferring category switches from the given data could be methodologically advanced. Whereas previous research relied on cost-intensive human scoring procedures (e.g., Troyer et al., 1997) of categories (i.e., farm animals vs. pets)—

which should be viewed as problematic, as the categories used by one person are often not congruent to the categories used by another person—the use of semantic distances for scoring verbal retrieval data might provide a viable computerized alternative (e.g., Kenett, 2019). The use of semantic distances as a measure of category switches could capitalize on the fact that representatives from a given subcategory (e.g., mammals) should be more closely related than representatives from two different subcategories (e.g., mammals vs. reptiles). Whenever a new semantic category is reached by participants, this should be depicted in a kind of bump in the semantic distance from one word to the other (e.g., Hills et al., 2015; Zemla & Austerweil, 2018).

Fourth, as typical Gr tests (i.e., fluency tests) are considered to be divergent in nature, that is, the task requirements are understood to be relatively unstructured because a variety of responses might be produced by test takers, there is no single veridical response, as opposed to classical tests of maximal performance (Cronbach, 1949). Due to that, typical Gr indicators are considered to be tests of *divergent thinking* (Guilford, 1967), a term that is also often used in the literature regarding creativity (e.g., Weiss et al., 2020). Further, established intelligence models specify creativity as a part of Gr. This placement should be scrutinized in future studies as, so far, there are only few multivariate studies examining the organization of general retrieval abilities and creative originality and creative fluency (e.g., Silvia et al., 2013).

Conclusion

Taken together, we applied a comprehensive battery of Gr tasks and compared competing measurement models. We showed that a higher-order factor model including first-order factors of ideational, word, expressional, and figural fluency provided good fit to the data, independent of different retrieval times in the verbal task domain. Predicting this higher-order factor of Gr with relevant cognitive covariates, we found that Gr still poses a strong and discriminant factor of cognitive ability. With that, our study provides insights into the dimensionality of Gr and extends our understanding of explanatory variables of individual differences in Gr.

References

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence*, *22*(2), 227–257. https://doi.org/10.1016/S0160-2896(96)90016-1

Amunts, J., Camilleri, J. A., Eickhoff, S. B., Heim, S., & Weis, S. (2020). Executive functions predict verbal fluency scores in healthy participants. *Scientific Reports*, *10*(1), 11141. https://doi.org/10.1038/s41598-020-65525-9

Ardila, A., Galeano, L. M., & Rosselli, M. (1998). Toward a Model of Neuropsychological Activity. *Neuropsychology Review*, *8*(4), 171–190. https://doi.org/10.1023/A:1021618218943

Ardila, A., Ostrosky-Solis, F., Rosselli, M., & Gómez, C. (2000). Age-related cognitive decline during normal aging: The complex effect of education. *Archives of Clinical Neuropsychology*, *15*(6), 495–513. https://doi.org/10.1093/arclin/15.6.495

Azuma, T. (2004). Working Memory and Perseveration in Verbal Fluency. *Neuropsychology*, *18*(1), 69–77. https://doi.org/10.1037/0894-4105.18.1.69

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, *63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baltes, P. B., Staudinger, U. M., & Lindenberger, U. (1999). Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology*, *50*(1), 471–507. https://doi.org/10.1146/annurev.psych.50.1.471

Beauducel, A., & Herzberg, P. Y. (2006). On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA.

Structural Equation Modeling: A Multidisciplinary Journal, *13*(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology*, *31*.

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*(4), 796–846. https://doi.org/10.1111/j.1467-6494.2011.00749.x

Bryan, J., Luszcz, M. A., & Crawford, J. R. (1997). Verbal Knowledge and Speed of Information Processing as Mediators of Age Differences in Verbal Fluency Performance Among Older Adults. *Psychology and Aging*, *12*(3), 473–478.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. Elsevier Science Publishers B.V.

Cohen, R. L. (1984). Individual differences in event memory: A case for nonstrategic factors. *Memory & Cognition*, *12*(6), 633–641. https://doi.org/10.3758/BF03213352

Cole, D. A., Perkins, C. E., & Zelkowitz, R. L. (2016). Impact of homogeneous and heterogeneous parceling strategies when latent variables represent multidimensional constructs. *Psychological Methods*, *21*(2), 164–174. https://doi.org/10.1037/met0000047

Collins, A. M., & Loftus, E. F. (1975). A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Conway, A. R. A., Jarrold, C., Kane, M. J., Miyake, A., & Towse, J. N. (2008). Variation in Working Memory: An Introduction. In A. R. A. Conway, C. Jarrold, & M. J. Kane (Eds.), *Variation in working memory*. Oxford University Press.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), Article 5. https://doi.org/10.3758/BF03196772

Cowan, N. (2005). *Working Memory Capacity (Essays in Cognitive Psychology)*. Tayler & Francis Ltd.

Cronbach, L. J. (1949). *Essentials of Psychological Testing*. Harper & Brothers, Publishers.

Danthiir, V., Wilhelm, O., & Schacht, A. (2005). Decision speed in intelligence tasks: Correctly an ability? *Psychology Science*, *47*(2), 200.

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor Models for Predicting Criteria by General and Specific Factors: Problems of Nonidentifiability and Alternative Solutions. *Journal of Intelligence*, *6*(3), 42. https://doi.org/10.3390/jintelligence6030042

Enders, C. K. (2010). *Applied missing data analysis.* Guilford Press.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*(1), 19–23.

Eysenck. (1987). Speed of information processing, reaction time, and the theory of intelligence. In P. A. Vernon (Ed.), *Speed of information-processing and intelligence* (pp. 21–67). Ablex Publishing.

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Forthmann, B., Goecke, B., & Beaty, R. (2022). *Planning Missing Data Designs for Human Ratings in Creativity Research: A Practical Guide* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/qbrhk

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing Speed as a Confounding Variable and the Measurement of Quality in Divergent Thinking. *Creativity Research Journal*, *29*(3), 257–269. https://doi.org/10.1080/10400419.2017.1360059

Forthmann, B., Jendryczko, D., Scharfen, J., Kleinkorres, R., Benedek, M., & Holling, H. (2019). Creative ideation, broad retrieval ability, and processing speed: A

confirmatory study of nested cognitive abilities. *Intelligence*, *75*, 59–72. https://doi.org/10.1016/j.intell.2019.04.006

Frischkorn, G. T., Wilhelm, O., & Oberauer, K. (2022). Process-oriented intelligence research: A review from the cognitive perspective. *Intelligence*, *94*, 1–14. https://doi.org/10.1016/j.intell.2022.101681

Goecke, B., Schmitz, F., & Wilhelm, O. (2021). Binding Costs in Processing Efficiency as Determinants of Cognitive Ability. *Journal of Intelligence*, *9*(2), 18. https://doi.org/10.3390/jintelligence9020018

Goecke, B., Staab, M., Schittenhelm, C., & Wilhelm, O. (2022). Stop Worrying about Multiple-Choice: Fact Knowledge Does Not Change with Response Format. *Journal of Intelligence*, *10*(4), 1–22. https://doi.org/10.3390/jintelligence10040102

Gordon, J. K., & Kindred, N. K. (2011). Word retrieval in ageing: An exploration of the task constraint hypothesis. *Aphasiology*, *25*(6–7), 774–788. https://doi.org/10.1080/02687038.2010.539699

Gruenewald, P. J., & Lockhead, G. R. (1980). The Free Recall of Category Examples. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(3), 225–240. https://doi.org/10.1037/0278-7393.6.3.225

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill, Inc.

Hakstian, A. R., & Cattell, R. B. (1978). Higher-Stratum Ability Structures on a Basis of Twenty Primary Abilities. *Journal of Educational Psychology*, *70*(5), 657–669. https://doi.org/10.1037/0022-0663.70.5.657

Hartshorne, J. K., & Germine, L. T. (2015). When Does Cognitive Functioning Peak? The Asynchronous Rise and Fall of Different Cognitive Abilities Across the Life Span. *Psychological Science*, *26*(4), 433–443.

Hedden, T., Lautenschlager, G., & Park, D. C. (2005). Contributions of Processing Ability and Knowledge to Verbal Memory Tasks across the Adult Life-Span. *The Quarterly Journal of Experimental Psychology Section A*, *58*(1), 169–190. https://doi.org/10.1080/02724980443000179

Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in Semantic Fields: How We Search Through Memory. *Topics in Cognitive Science*, *7*(3), 513–534. https://doi.org/10.1111/tops.12151

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Inquisit 6 (6.5.2). (2022). Millisecond Software. https://www.millisecond.com

Jäger, A. O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M., & Beauducel, A. (2006). *Berliner Intelligenzstrukturtest für Jugendliche: Begabungs- und Hochbegabungsdiagnostik [Berlin–Structure–Intelligence Test: Diagnosis of Giftedness and High Ability]*. Hogrefe. https://www.testzentrale.de/shop/berliner-intelligenzstrukturtest-fuer-jugendliche-begabungs-und-hochbegabungsdiagnostik.html

Jensen. (2006). *Clocking the mind: Mental chronometry and individual differences*. Elsevier.

Jewsbury, P. A., & Bowden, S. C. (2017). Construct Validity of Fluency and Implications for the Factorial Structure of Memory. *Journal of Psychoeducational Assessment*, *35*(5), 460–481. https://doi.org/10.1177/0734282916648041

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working Memory Capacity and Fluid Intelligence Are Strongly Related Constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 66–71. https://doi.org/10.1037/0033-2909.131.1.66

Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences*, *27*, 11–16. https://doi.org/10.1016/j.cobeha.2018.08.010

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin*, *136*(5), 849–874. https://doi.org/10.1037/a0019842

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence*, *14*(4), 389–433. https://doi.org/10.1016/S0160-2896(05)80012-1

Lehrl, S. (2005). *Mehrfachwahl-Wortschatz-Intelligenztest [Multiple-Choice Vocabulary Intelligence Test]*. Spitta.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits. *Structural Equation*

Modeling: A Multidisciplinary Journal, *9*(2), 151–173. https://doi.org/10.1207/S15328007SEM0902_1

Mayr, U., & Kliegl, R. (2000). Task-set switching and long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1124–1140. https://doi.org/10.1037/0278-7393.26.5.1124

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

McGrew, K. S., Schneider, W. J., Decker, S. L., & Bulut, O. (2023). A Psychometric Network Analysis of CHC Intelligence Measures: Implications for Research, Theory, and Interpretation of Broad CHC Scores "Beyond g." *Journal of Intelligence*, *11*(1), 1–33. https://doi.org/10.3390/jintelligence11010019

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Metternich, B., Buschmann, F., Wagner, K., Schulze-Bonhage, A., & Kriston, L. (2014). Verbal Fluency in Focal Epilepsy: A Systematic Review and Meta-analysis. *Neuropsychology Review*, *24*(2), 200–218. https://doi.org/10.1007/s11065-014-9255-8

Meyer, D., & Schvaneveldt, R. (1971). Facilitation in Recognizing Pairs of Words: Evidence of a Dependence Between Retrieval Operations. *Journal of Experimental Psychology*, *90*(2), 227–234. https://doi.org/10.1037/h0031564

Miroshnik, K. G., Forthmann, B., Karwowski, M., & Benedek, M. (2023). The relationship of divergent thinking with broad retrieval ability and processing speed: A meta-analysis. *Intelligence*, *98*, 101739. https://doi.org/10.1016/j.intell.2023.101739

Mogle, J. A., Lovett, B. J., Stawski, R. S., & Sliwinski, M. J. (2008). What's so special about working memory? An examination of the relationships among working memory, secondary memory, and fluid intelligence. *Psychological Science*, *19*(11), 1071–1077.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working Memory and Intelligence--Their Correlation and Their Relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), Article 1. https://doi.org/10.1037/0033-2909.131.1.61

Perret, E. (1974). The left frontal lobe of man and the suppression of habitual responses in verbal categorical behaviour☆. *Neuropsychologia*, *12*(3), 323–330. https://doi.org/10.1016/0028-3932(74)90047-5

Pornprasertmanit, S., Miller, P., Schoemann, A., & Jorgensen, T. D. (2021). *Simsem: SIMulated Structural Equation Modeling. R package.* (0.5-16). https://CRAN.Rproject.org/package=simsem

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raykov, T., & Marcoulides, G. A. (2011). Classical Item Analysis Using Latent Variable Modeling: A Note on a Direct Evaluation Procedure. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(2), 315–324. https://doi.org/10.1080/10705511.2011.557347

Robinson, G., Shallice, T., Bozzali, M., & Cipolotti, L. (2012). The differing roles of the frontal cortex in fluency tests. *Brain*, *135*(7), 2202–2214. https://doi.org/10.1093/brain/aws142

Rogers, R. D., & Monsell, S. (1995). Costs of a Predictable Switch Between Simple Cognitive Tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–231.

Rosen, V. M., & Engle, R. W. (1997). The Role of Working Memory Capacity in Retrieval. *Journal of Experimental Psychology: General*, *126*(3), 211–227. https://doi.org/10.1037/0096-3445.126.3.211

Rosseel, Y. (2012). **lavaan**: An *R* Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Sauzéon, H., Raboutet, C., Rodrigues, J., Langevin, S., Schelstraete, M. A., Feyereisen, P., Hupet, M., & N'Kaoua, B. (2011). Verbal Knowledge as a Compensation Determinant of Adult Age Differences in Verbal Fluency Tasks over Time. *Journal of Adult Development*, *18*(3), 144–154. https://doi.org/10.1007/s10804-010-9107-6

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037//1082-989X.7.2.147

Schelble, J. L., Therriault, D. J., & Miller, M. D. (2012). Classifying retrieval strategies as a function of working memory. *Memory & Cognition*, *40*(2), 218–230. https://doi.org/10.3758/s13421-011-0149-1

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). 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

Schmidt, K. H., & Metzler, P. (1992). *Wortschatztest [Vocabulary Test]*. Beltz.

Schmitz, F., & Wilhelm, O. (2016). Modeling Mental Speed: Decomposing Response Time Distributions in Elementary Cognitive Tasks and Correlations with Working

Memory Capacity and Fluid Intelligence. *Journal of Intelligence*, *4*(4), Article 4. https://doi.org/10.3390/jintelligence4040013

Schneider, W. J., & McGrew, K. S. (2018). The Cattell–Horn–Carroll Theory of Cognitive Abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 73–163). The Guilford Press.

Schroeders, U. (2018, September). *Entwicklung und Validierung eines korpusbasierten Wortschatztests [Development and Validation of a Corpus-Based Vocabulary Test]*. DPGS [German Psychological Association], Frankfurt am Main, Germany.

Shah, P., Miyake, A., Miyake, A., & Shah, P. (1999). Models of Working Memory: An Introduction. In *Models of Working Memory: Mechanisms of active maintenance and executive control.* (pp. 1–27). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.004

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00772

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working Memory Capacity and Fluid Intelligence: Maintenance and Disengagement. *Perspectives on Psychological Science*, *11*(6), 771–799. https://doi.org/10.1177/1745691616650647

Silvia, P. J., Beaty, R. E., & Nusbaum, E. C. (2013). Verbal fluency and creativity: General and specific contributions of broad retrieval ability (Gr) factors to divergent thinking. *Intelligence*, *41*(5), 328–340. https://doi.org/10.1016/j.intell.2013.05.004

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*(3), 214–241. https://doi.org/10.1037/h0036351

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

Thomsen, D. K. (2015). Autobiographical Periods: A Review and Central Components of a Theory. *Review of General Psychology*, *19*(3), 294–310. https://doi.org/10.1037/gpr0000043

Thurstone, L. L. (1938). *Primary Mental Abilities*. University of Chicago Press.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology*, *11*(1), 138–146. https://doi.org/10.1037/0894-4105.11.1.138

Tukey, J. W. (1977). *Exploratory Data Analysis* (2nd ed.). Addison-Wesley.

Tulving, E. (1972). Episodic and Semantic Memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory*. Academic Press.

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin*, *145*(1), 79–139. https://doi.org/10.1037/bul0000176

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Working memory capacity and retrieval from long-term memory: The role of controlled search. *Memory & Cognition*, *41*(2), 242–254. https://doi.org/10.3758/s13421-012-0261-x

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search

from secondary memory. *Psychological Review*, *114*(1), 104–132. https://doi.org/10.1037/0033-295X.114.1.104

Unsworth, N., Miller, J. D., Lakey, C. E., Young, D. L., Meeks, J. T., Campbell, W. K., & Goodie, A. S. (2009). Exploring the Relations Among Executive Functions, Fluid Intelligence, and Personality. *Journal of Individual Differences*, *30*(4), 194–200. https://doi.org/10.1027/1614-0001.30.4.194

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2011). Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *The Quarterly Journal of Experimental Psychology*, *64*(3), 447–466. https://doi.org/10.1080/17470218.2010.505292

Vannorsdall, T. D., Maroof, D. A., Gordon, B., & Schretlen, D. J. (2012). Ideational fluency as a domain of human cognition. *Neuropsychology*, *26*(3), 400–405. https://doi.org/10.1037/a0027989

von Stumm, S. (2013). Investment Traits and Intelligence in Adulthood: Assessment and Associations. *Journal of Individual Differences*, *34*(2), 82–89. https://doi.org/10.1027/1614-0001/a000101

Vurgun, R. (2014). *Subliminales syntaktisches Priming: Einfluss von maskierten Pronomen auf die Erkennung von sichtbaren Verben in lexikalischen Entscheidungsaufgaben [Subliminal syntactic priming: Influence of masked pronouns on the recognition of visible verbs in lexical decision tasks]* [Unpubslished Diploma Thesis]. https://utheses.univie.ac.at/detail/30102#

Weiss, S., Steger, D., Kaur, Y., Hildebrandt, A., Schroeders, U., & Wilhelm, O. (2020). On the Trail of Creativity: Dimensionality of Divergent Thinking and Its Relation

With Cognitive Abilities, Personality, and Insight. *European Journal of Personality*, *35*, 22. https://doi.org/10.1002/per.2288

Whiteside, D. M., Kealey, T., Semla, M., Luu, H., Rice, L., Basso, M. R., & Roper, B. (2016). Verbal Fluency: Language or Executive Function Measure? *Applied Neuropsychology: Adult*, *23*(1), 29–34. https://doi.org/10.1080/23279095.2015.1004574

Wilhelm, O. (2005). Measuring Reasoning Ability. In *Handbook of Understanding and Measuring Intelligence* (pp. 373–392). SAGE Publications, Inc. https://doi.org/10.4135/9781452233529.n21

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00433

Wilhelm, O., & Kyllonen, P. (2021). To predict the future, consider the past: Revisiting Carroll (1993) as a guide to the future of intelligence research. *Intelligence*, *89*, 101585. https://doi.org/10.1016/j.intell.2021.101585

Wilhelm, O., & Schroeders, U. (2019). Intelligence. In R. Sternberg & J. Funke (Eds.), *The Psychology of Human Thought: An Introduction* (pp. 257–277). Heidelberg University Publishing.

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, *1*(1), 89–106. https://doi.org/10.3758/BF03200763

Zemla, J. C., & Austerweil, J. L. (2018). Estimating Semantic Networks of Groups and Individuals from Fluency Data. *Computational Brain & Behavior*, *1*(1), 36–58. https://doi.org/10.1007/s42113-018-0003-7

Ziegler, M., Danay, E., Heene, M., Asendorpf, J., & Bühner, M. (2012). Openness, fluid intelligence, and crystallized intelligence: Toward an integrative model. *Journal of Research in Personality*, *46*(2), 173–183. https://doi.org/10.1016/j.jrp.2012.01.002

VI Epilogue

Summary

Taken together, the manuscripts of this dissertation addressed some specific of the many interesting thus far unanswered research questions regarding individual differences in declarative knowledge and related constructs. With that, this dissertation helps to put declarative knowledge into perspective regarding its measurement and its linkage to other variables. At the same time, it lays the foundation for important future work. In the first and second manuscripts, I focused on studying overclaiming tendencies as a potential bias of self-reported knowledge. Within the third manuscript, I investigated possible psychometric consequences of using competing response formats for measuring individual differences in declarative knowledge. With the fourth manuscript, I provide the empirical results of a large multivariate study, including both measures of crystallized intelligence and broad retrieval ability, which informs us not only about their crucial relationship but also about their nomological net.

In the following sections, I will summarize both the theoretical contributions and empirical findings of the four manuscripts comprising this cumulative dissertation before moving on to a broader discussion. Lastly, I will provide suggestions for future research that might aid the study of knowledge and related constructs above and beyond what is presented in the manuscripts of this dissertation.

Manuscript I: Testing Competing Claims About Overclaiming

The first manuscript (Goecke et al., 2020) helps us to broaden our understanding regarding self-reports of knowledge as a potential measurement instrument of declarative knowledge. In a multivariate study with adults (see also Weiss et al., 2020, N = 298), I investigated self-reported knowledge as measured by Overclaiming Questionnaires (OCQs, c.f., Paulhus et al., 2003) and, in the same vein, examined how so-called overclaiming tendencies would distort the self-reported knowledge levels of participants (e.g., Hülür et al., 2011, 2014). Moreover, I examined the nomological net of both self-reported knowledge and overclaiming by deriving hypotheses based on several competing theoretical perspectives of

why overclaiming behavior can be observed. From these perspectives, overclaiming on OCQs is either seen as a self-enhancement tendency (e.g. Paulhus et al., 2003), as a product of a cognitive processing bias (e.g., Müller & Moshagen, 2018), as a potential index of declarative knowledge (Paulhus & Harms, 2004), or as a type of creative engagement (e.g., Dunlop et al., 2017). Therefore, based on the mentioned perspectives, we added covariates from both the ability (e.g., fluid intelligence, declarative knowledge, divergent thinking [i.e., creativity including fluency and originality], and faking ability) and the personality (honesty-humility and the creativity facet of openness to new experiences) framework to our study. Inspired by previous work (e.g., Bing et al., 2011; Fell et al., 2017; Hülür et al., 2011, 2014; Phillips & Clancy, 1972; Randall & Fernandes, 1991), the introduction of the first manuscript also includes a critical view of previously recommended scoring approaches regarding OCQs (i.e., analytical approaches based on signal detection theory) and provides a more straightforward and parsimonious alternative based on the framework of confirmatory factor analysis. Along with the manuscript, we published a newly developed OCQ (including reals and foils) in German language that covers five broad knowledge domains (natural sciences, humanities, life sciences, social sciences, and pop culture), which correspond to previously identified domains of declarative knowledge (c.f., Steger et al., 2019) and thus integrate the here administered self-reported knowledge measure into the research on declarative knowledge.

The study provides evidence for a latent variable of self-reported knowledge being severely inflated by overclaiming, with 50.5% of the variance in self-reported knowledge explained by overclaiming alone. With that, the results replicate earlier findings (e.g., Hülür et al., 2011, 2014) and provide further support for the assumption that self-reported knowledge might not be a pertinent measure of actual levels of knowledge. This is also reflected in the moderate regression weight of a latent variable of declarative knowledge towards self-reported knowledge. Regarding overclaiming, the results indicate that overclaiming should not be merely taken as a tendency to self-enhance (Paulhus, 2011), because it was neither predicted by honesty-humility nor by faking ability (c.f., Geiger et al., 2018). Further, the evidence for overclaiming being driven by creative personality

characteristics such as the creativity facet of openness (c.f., Dunlop et al., 2017) or creativity in the form of divergent thinking (i.e., creativity including fluency and originality) was weak. However, divergent thinking was predictive of self-reported knowledge. This might, in part, be explained through the role that divergent thinking plays as a sub-facet of broad retrieval ability: both depict the ability retrieve information from long-term memory fluently, and this refers to the argument that retrieval abilities should be somewhat crucial for individual differences in knowledge. After all, it cannot be negated that self-reported knowledge has at least a little to do with actual levels of declarative knowledge, although measuring knowledge by means of self-reports might not be the most valid method.

Manuscript II: The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children

Addressing the nomological net of self-reported knowledge, overclaiming, and declarative knowledge was also the focus of the second manuscript (Goecke, Schroeders, et al., 2022) of this dissertation. In fact, this large multivariate study (N = 897) builds on the theoretical ideas of the first manuscript (Goecke et al., 2020) and transfers the previously derived theoretical perspectives on overclaiming as a severe distortion of self-reported knowledge to a sample of children (i.e., a hitherto understudied population). The original idea of measuring overclaiming (although labeled differently at that time) was formulated almost a century ago (c.f., Raubenheimer, 1925) and studied in a sample of young children. However, since then, there has existed an apparent lack of studies (for exceptions, see Butler & Nelson, 2021; PISA 2012, OECD, 2013; Raubenheimer, 1925) investigating overclaiming and self-reported knowledge in young children. Our study adds important evidence to this existing body of literature, as there might be unforeseen differences in both overclaiming behavior and self-reported knowledge in children that can be attributed to ongoing meta-cognitive development in children (e.g., W. Schneider, 2008). For example, whether self-reported knowledge would be as inflated by overclaiming in children as it was found to be in adults (Goecke et al., 2020; Hülür et al., 2011, 2014) was an open question. If this were not the case, this would have consequences for studying developmental trajectories of knowledge in

children and developmental trajectories of overclaiming. Theoretical accounts of this change could be an intriguing field of research. Based on the derived perspectives on overclaiming in adults, we again added covariates from both the ability (e.g., fluid intelligence, declarative knowledge, divergent thinking [i.e., fluency]) and the personality (parent reports of honesty-humility, openness, and self-reports of epistemic curiosity) framework to the study presented in manuscript II. Just like in the first manuscript, we published a newly developed OCQ (including reals and foils) in German language tailored to children from third grade that covers three broad knowledge domains (natural sciences, humanities, social sciences), which reflect the school curriculum (c.f., Schroeders et al., 2016, 2020).

The empirical results of this study are mostly in line with the evidence regarding self-reported knowledge and overclaiming in adults (c.f., Goecke et al., 2020). Self-reported knowledge in children is even slightly stronger predicted by overclaiming than in adults (62.3% vs. 50.5%, respectively), which thus once more calls into question the utility of self-reports of knowledge as being valid measurement instruments of actual knowledge. This is again emphasized through an only small increment of declarative knowledge predicting self-reported knowledge beyond overclaiming. In addition to that, we found no relationship between overclaiming with any of the personality-related variables (honesty-humility, openness, and epistemic curiosity), which somewhat challenges the theoretical perspectives of overclaiming being indicative of either self-enhancement tendencies (e.g., Paulhus et al., 2003) or some form of open personality (e.g., Dunlop et al., 2017). However, openness was found to have low but substantial predictive power for self-reported knowledge, which is in line with literature that poses openness as an important predictor of knowledge acquisition (e.g., Ackerman, 1996; von Stumm, 2013; Ziegler et al., 2012). In contrast to that, it was surprising that fluency (a measure of broad retrieval ability [Gr]) was predictive of overclaiming but not of self-reported knowledge. With that, the results of the second manuscript strengthen the evidence found in the first manuscript. Moreover, the results inform us that overclaiming and self-reported knowledge in children are not that different from the same variables in adults. Taken together, self-reported knowledge has to be viewed

critically because such knowledge estimates are severely biased by overclaiming. In turn, overclaiming as measured by mean familiarity ratings on OCQs is a stable construct of individual differences, but the nomological network of overclaiming is hard to grasp, and further attempts to study overclaiming should involve alternative measurement instruments. If we want to understand why self-reported knowledge is inflated by overclaiming, we will not be able to avoid variations of test instruments assessing both self-reported knowledge and overclaiming in future research.

Manuscript III: Stop Worrying About Multiple-Choice: Knowledge Does Not Change with Response Format

Within the third manuscript (Goecke, Staab, et al., 2022), I studied the potential psychometric consequences of applying three competing response formats (multiple-choice, cued open-ended, and open-ended formats) for the assessment of declarative knowledge across two online studies (N_{Study1} = 142, N_{Study2} = 300). As reservations regarding multiple-choice formats are frequently brought forward in both the literature (e.g., Becker & Johnston, 1999; Hickson & Reed, 2011; Krieg & Uyar, 2001) and in applied settings (e.g., in conversations with university students), I aimed at investigating whether declarative knowledge as understood from a maximal performance perspective (Cronbach, 1949) changes as a function of the applied response format. If such a response format induced change was the case, this would have serious implications regarding the debate of how knowledge is best be measured, as it might indicate that different cognitive processes (e.g., recognition vs. recall; c.f., Tulving & Watkins, 1972) are tapped by different response formats. Thus, I formulated three plausible psychometric scenarios that might be observable when different response formats are used to measure the supposedly same underlying construct (i.e., declarative knowledge) and put them to a critical test: first, the scenario that neither the mean structure nor the covariance structure will be affected by different response formats; second, the scenario that the mean structure is affected by different response formats, but the covariance structure is not; and lastly, that both the mean and the covariance structure is affected by applying different response formats. Within both studies, we used a newly compiled item set (72 items) of four

broad knowledge domains (natural sciences, humanities, social sciences, and life sciences) derived from earlier work (Steger et al., 2019) and implemented a control mechanism for the unproctored study design to control for cheating behavior (c.f., Steger et al., 2021).

The results of the two studies provide evidence for effects of the response format on the mean structure of the applied declarative knowledge test but indicate that the covariance structure remains largely unaffected. The mean effects were visible in both plain descriptive statistics of the applied tests (i.e., effect sizes ranged between d = .33 - 1.49) and more formal logistic regression analysis using single responses as the criterion variable. In contrast to that, latent variable analyses revealed that modeling factors indicated by single response formats yielded latent correlations at unity and that declarative knowledge is better modeled according to its underlying broad knowledge domains, with a higher-order factor articulated at the apex of the hierarchy. This result was further strengthened by the failed attempt to model nested response format factors (which were not reliable and exhibited no significant variances). Therefore, the here-presented data provide no indication that different underlying cognitive processes (e.g., recognition vs. recall), are at work when responding to declarative knowledge questions. Further, the results also replicated earlier research modeling declarative knowledge (e.g., Steger et al., 2019) as a hierarchical construct. Taken together, the studies of manuscript III provide meaningful empirical evidence for a lack of different cognitive processes being activated while responding to knowledge items in different response formats.

Manuscript IV: Driving Factors of Individual Differences in Broad Retrieval Ability:* Gr *Is More Than the Sum of Its Parts

In the fourth manuscript of this dissertation (Goecke et al., 2023) I studied broad retrieval ability (Gr) by means of an extensive multivariate study (N = 331). In this study, I examined whether specific retrieval affordances, such as differing time periods (e.g., short vs. long periods of retrieval) as imposed on verbal retrieval tasks, could be psychometrically distinguished from more general task affordances. In addition to that, I tested whether broad retrieval ability (c.f., Carroll, 1993; Schneider & McGrew, 2018) must be understood as a mere

linear function of other related ability constructs or as more than the sum of its parts (c.f., Cronbach & Meehl, 1955). To this end, both a broad measurement of Gr, including several more specific abilities (ideational fluency, word fluency, expressional fluency, and figural fluency), and also a number of relevant covariates were included in the study. The covariates were introduced based on theory-driven considerations regarding the presumed cognitive processes leading to individual differences in Gr based on a four-component model of retrieval ability (Rosen & Engle, 1997). Importantly, these components could be mapped onto working memory capacity, secondary memory, mental speed, and crystallized intelligence (including declarative knowledge and vocabulary).

The results of this study indicate that no specific retrieval affordances regarding time periods can be distinguished psychometrically from the general ability to retrieve information from long-term memory. This is reflected by the finding that verbal Gr is best modeled according to its subordinate abilities, that is, as a higher-order model with Gr at the top of the hierarchy and factors of ideational, word, expressional, and figural fluency modeled as first-order factors underneath. Therefore, our study replicates previous findings modeling Gr as a higher-order factor model (e.g., Silvia et al., 2013) and further extends this previous research by including a more diverse set of retrieval tasks. In addition to that, the study adds substantially to the existing body of work regarding Gr by examining a large nomological net around Gr. We specifically tested whether the administered covariates showed incremental predictive validity towards Gr above and beyond their shared variances. To this end, I proposed a bifactorial measurement model of the cognitive covariates that controlled for their shared variance by computing a general factor (*g*) accounting for the communality of all indicators of the covariates. We identified a strong and reliable general factor marked through indicators of working memory capacity, but also reliable nested factors of mental speed, secondary memory, and Gc (including a measure of vocabulary, a multiple-choice declarative knowledge test, and an open-ended declarative knowledge test). On a related note, this measurement model replicated the finding that there is no effect of response format on the covariance structure of declarative knowledge tests, as presented in manuscript III (Goecke

et al., 2022). The combination of the measurement model of Gr, on the one hand, and the measurement model of the covariates, on the other hand, resulted in a structural equation model where the higher-order factor of Gr was predicted by all covariates, including *g*. Taken together, we found that 62% of the variance in Gr was explained, with Gc being the second strongest predictor of Gr after *g*. In sum, the results of this study endorse the interpretation of Gr being a strong and distinct factor of cognitive abilities, but they also emphasize the vital role that (relevant prior) knowledge (i.e., Gc or crystallized information) plays for Gr. Considering the arguments regarding Gc made in both manuscript IV (Goecke et al., 2023) and in the prologue of this dissertation, the same could apply vice versa.

Looking Towards the Empirical Findings and Beyond

Self-reported knowledge hardly serves as a good proxy for individual differences in declarative knowledge, as it is severely biased by overclaiming; in turn, overclaiming is only hard to grasp from a nomological network perspective (manuscripts I and II). More reliable measurement instruments of declarative knowledge rely on measures of maximal performance, whereas the administered response formats are not important for assessing individual differences (manuscript III). Accessing and retrieving previously learned knowledge from long-term memory is arguably akin to prototypical task requirements as imposed by tasks tapping broad retrieval ability, and thus investigating the nomological net of broad retrieval ability broadens our understanding of the cognitive underpinnings involved in knowledge retrieval (manuscript IV). In the following sections, I will focus on the questions initially proposed in this dissertation and derive recommendations for the measurement of declarative knowledge.

Question 1: How Is Self-Reported Knowledge to Be Placed Within the Existing Ability Structure?

How much simpler would our world be if individuals were capable of providing accurate estimates of their own abilities? Although such utopian thought experiments might be interesting, the empirical data speak plainly to us (e.g., Freund & Kasten, 2012; Mabe &

West, 1982; Zell & Krizan, 2014). In line with previous literature, I provided further evidence that self-reported knowledge has only little to do with actual declarative knowledge. Therefore, using self-reported knowledge as a proxy of actual knowledge should be continuously problematized.

Although the relations between self-reported knowledge and declarative knowledge inferred from the current datasets are neither approaching zero nor unity (c.f., manuscripts I and II) and might, therefore, also be interpreted as a glass half full (i.e., reflecting at least some real fact knowledge), self-reported knowledge should not be used as a substitute for real measures of ability. In fact, I argue that no self-report instrument whatsoever should be treated as an adequate substitute for any cognitive ability test. This argument is supported by other research areas (e.g. the measurement of socioemotional constructs) where it was found that the application of self-reports and ability tests leads to the assessment of different constructs (Olderbak & Wilhelm, 2020). In line with this argument, self-reported knowledge cannot be considered a measure of cognitive ability for several reasons because these test instruments do not meet the standards imposed on actual tests of maximal performance (e.g., Cronbach, 1949): First, self-reports lack the principal instruction to show maximal performance. This is problematic because in order to assess an ability like declarative knowledge, test takers should be instructed to show their best performance in a proctored test setting. Where one person might outperform others, it is important to make expectations transparent. Second, the lack of veridicality of self-reports of knowledge (e.g., Likert scales ranging from 1 for little to 5 for much knowledge regarding a target item) prevents an objective evaluation of performance. This is problematic because participants might have a different understanding of what *familiarity* with an item (e.g., a term or word, or sentence) actually means and thus might interpret and use the response scale differently (e.g., Ackerman & Ellingsen, 2014; Weiss & Roberts, 2018). Third, knowledge purporting in the guise of overclaiming biases estimates derived from self-reported knowledge questionnaires, which contradicts a basal requirement of performance tests. Namely the requirements of test data (c.f., T-data; Cattell, 1958) of being immune to various biases, such as faking good, social

desirability, acquiescence, or malingering. Thus, placing self-reports of knowledge in the existing ability structure is steadily accompanied by an overhanging validity restriction that poses the question of what scores on self-reports of knowledge might actually represent when they are not representative of actual declarative knowledge.

One possible response regarding the question of what self-reported knowledge scores represent was articulated in a study presented by Rolfhus and Ackerman (1996). In this study, the authors found substantial correlations between their measure of self-reported knowledge and a six-dimensional interest scale (c.f., RIASEC, Holland, 1977). They concluded that "as predicted, interests and knowledge scales sharing the same content were correlated, in some cases quite strongly" (Rolfhus & Ackerman, 1996, p. 184) and that self-reported knowledge might be more reflective of personal interests than actual knowledge "although a causal attribution cannot be made" (Rolfhus & Ackerman, 1996, p. 184). This is also in line with related literature arguing that self-reports of cognitive abilities might be more related to personality characteristics than to cognitive functioning (Herreen & Zajac, 2017). This is interesting because these results provide insights into how self-reported knowledge scales might be used meaningfully, namely as a measure of the domain-specific interests of test takers.

In addition to that, variations of self-reported knowledge scales might also be used to inform test takers about their inaccuracies regarding their own knowledge. For instance, in another study (Ackerman & Ellingsen, 2014), participants were presented with lists of words and asked to state whether they were able to define these words. After stating for every word stimulus whether they were able to do so, whether they were not able to do so, or whether they were unsure, participants were presented again with the initial word list and asked to provide actual definitions. Self-reports of knowledge might be used in a similar way if they were extended by a second stage, where test takers are actually asked to provide definitions regarding the words they claimed to know. Surely not all test takers would be able to provide the correct definitions of the words they claimed to know. The inherent feedback provided through that second stage (i.e., did I really know something, or did I just claim to know it?)

might be informative for test takers in two ways: first, they would arguably learn where they lack knowledge; and second, they might be informed regarding their possible interests through being informed about the topics they wish they had known. This way, self-reports of knowledge could, for example, be used to complement so-called online self-assessments (e.g., Reiss et al., 2009) that strive to inform prospective students regarding their own abilities, motivations, competencies, and interests.

Despite an abundance of validated maximal performance measures and a large body of literature regarding the over- and underestimation of cognitive abilities (for example, in light of the Dunning-Kruger effect [Kruger & Dunning, 1999], which has also been reviewed critically [c.f., Gignac & Zajenkowski, 2020; Hofer et al., 2022]), self-reports of abilities or knowledge are sometimes applied to infer knowledge levels (e.g., Cotter et al., 2023). As stated above, using self-reports to measure what should be understood as an ability in the first place can also be observed in other areas (e.g., the literature on emotional intelligence; Olderbak & Wilhelm, 2020). Therefore, I would like to suggest a possible bridge between self-reports of knowledge (Q-data) and biographical self-reports (L-data) (c.f., Cattell, 1958) in order to provide an interesting alternative measurement format for approximating actual knowledge levels in a pretzel-like twist.

Considering that all declarative knowledge might be the result of an individual's biography (e.g., Cattell, 1987), researchers could capitalize on the evaluable biographical information of individuals to infer their knowledge levels. For example, L-data (i.e., biographical data) were used by NASA to identify scientists and engineers (Institute for Behavioral Research in Creativity, 1968). Such use makes sense as, for example, the likelihood of an individual knowing who Antoni Gaudí was might be considerably increased if said individual had visited Barcelona and the building sight of La Sagrada Familia in their life (although, of course, this is not guaranteed). In the same vein, this information (a visit to Barcelona) is verifiable and does not require too much introspection. As such, for a considerable number of declarative knowledge questions, a biographical question as a counterpart might be thinkable. Taking the idea of using self-reports as a valid measure of

knowledge a step further might therefore include the shift from self-report questionnaire data (e.g., beliefs assessed on a Likert scale) towards the use of more verifiable L-data (i.e., verifiable biographical questions) to approximate knowledge. In addition to previous ideas regarding investment traits driving knowledge acquisition, such an endeavor might also shed light on the idiosyncratic process of knowledge acquisition from a more biographical point of view.

Taken together, it might not be as much a question of how self-reports of knowledge are to be placed within the existing ability structure but more of how self-reports (of knowledge) might be meaningfully used to inform test takers about their own inaccuracies regarding their knowledge.

Question 2: How Do Different Response Formats Affect the Measurement of Declarative Knowledge?

Inferring latent constructs like declarative knowledge from behavioral measurements like fact knowledge questions capitalizes on the idea that the constructs of interest supposedly exist independent of the inquiry of measurement we apply (Cronbach & Meehl, 1955). That is, from a multi-trait-multi-method perspective (Campbell & Fiske, 1959), it is desirable for the measurement of psychological constructs when they are not subject to change once we change the instruments supposedly inferring them. In the case of declarative knowledge, this notion has several implications. Apart from the idea that any item of a test should be easily exchangeable (Spearman, 1904), which might be debatable in the case of declarative knowledge (Schroeders et al., 2021; Steger et al., 2022), this idea means it should be easily possible to switch from one, maybe more restricted response format, to another, maybe less restricted response format, without sacrificing any validity.

As presented in manuscript III, I found that although the difficulty of declarative knowledge items changes due to the response format, the response format is largely irrelevant in accounting for the covariance structure. Importantly, the results of the above-presented latent variable approach indicate that no meaningful variance regarding response formats is

left to be explained in the indicators after capturing the communality between indicators of different response formats and within knowledge domains. Specified method factors have no substantial variance. This also indicates that any other differential relations are impossible as only insufficient communality amongst the methods remains. Due to this set of results, a good response to the initially asked question should read that the measurement of individual differences in declarative knowledge is not affected by different response formats as long as all test takers are presented with the same test.

However, considering that knowledge might be expressed in other forms than simply responding to fact knowledge questions in multiple-choice or (cued) open-ended response formats requires thinking one step further. Naturally, the response formats used and proposed in the above-presented study (c.f., manuscript III, Goecke et al., 2022) must be considered a specific selection of possible response formats. Numerous other response and test formats are thinkable; and although the conclusions drawn from the data provided are warranted, it is important to question whether they would still hold once more even more complex response formats (e.g., writing an essay to a given prompt) are considered, which may involve more complex cognitive processes after all, as for instance articulated in Bloom's taxonomy (e.g., Krathwohl, 2002). On the one hand, it seems easily conceivable that responding to fact knowledge questions regarding, say, medicine--may they be presented in a multiple-choice format or in an open-ended response format (where the correct response still requires one or a small number of words)—might require different or less complicated retrieval processes than writing an essay about the state of medical ethics in light of the COVID-19 pandemic. On the other hand, using more complicated test formats (e.g., essays) as indicators of (fact) knowledge might only marginally affect the construct validity of what is to be assessed in the first place. Apart from that, scoring essays requires inherently more resources, and automated scoring procedures are currently still facing numerous challenges (Ramesh & Sanampudi, 2022).

If, as I proposed in the above-presented work—but as was also indicated in much earlier (c.f., Tulving & Watkins, 1973)—different response formats were to be placed on a

fictional response format continuum, the multiple-choice format as one of the most restrictive response formats would be located on one pole of the continuum, whereas less restricted response formats (e.g., open-ended response formats) "would cover the rest of the continuum in different manifestations, spanned from cued open-ended response formats that only require responding with one word to open-ended response formats that are akin to essays. Depending on the location of a response format on the continuum, the requirement of different cognitive processes for responding to an item-stem might change" (manuscript III, Goecke et al., 2022, p. 5). Thus, further variations of response formats for the assessment of knowledge should be critically tested and reviewed in future research in order to further elaborate on the question proposed in this dissertation. Until then, if two tests are perfectly correlated, the more economical (faster to administer) and more objective test (easier to score) should be preferred in real-life test settings. Hence, let's hear it for multiple-choice tests.

Question 3: Which Cognitive Abilities Explain Individual Differences In Broad Retrieval Ability?

The third overarching question of this dissertation aimed at probing our knowledge regarding Gc by studying Gr and its cognitive underpinnings. Presumed cognitive underpinnings contributing to differential levels in Gr were described in various component models that aim to explain how individual differences in verbal fluency emerge (e.g., Gruenewald & Lockhead, 1980; Rosen & Engle, 1997; Unsworth et al., 2011). Importantly, the four-component model proposed by Rosen & Engle (1997) arguably includes the main ingredients of differential levels in verbal fluency and, therefore can also be considered important for explaining differential levels in declarative knowledge. In manuscript IV of this dissertation, I derived hypotheses regarding these components and their possible instantiations in terms of cognitive abilities, including mental speed, working memory capacity, secondary memory, and crystallized intelligence (c.f., Rosen & Engle, 1997; Unsworth et al., 2011). In this regard, the here-presented study (manuscript IV, Goecke et al., 2023) draws an unequivocal picture: Gr is more than the sum of its parts, as it was not possible to fully delineate a higher-order factor of Gr through the other cognitive ability measures that

arguably drive individual differences in Gr. This suggests that Gr represents a unique and distinct construct that captures something beyond what is measured by other cognitive abilities. In addition to that, all of the included cognitive abilities, including a factor of Gc, showed incremental predictive validity towards Gr above and beyond a general factor (i.e., their communality). As studying the nomological net of Gr almost certainly also informs about Gc, it is central to consider the empirical findings of manuscript IV (Goecke et al., 2023) in light of the study of declarative knowledge. Why?

Consider the following two items, which might both be part of a test battery set out to measure declarative knowledge on the one hand and verbal fluency as an indicator of broad retrieval ability on the other hand, respectively: First, "Which rivers run through the city of Pforzheim?" and second, "Name as many rivers that run through Germany as you can!". Evidently, the number of (correct) responses to both items is overlapping. Whereas the declarative knowledge item requires test takers to find the correct information (i.e., the three rivers running through Pforzheim) amongst a multitude of irrelevant noise (i.e., all other rivers), the verbal fluency item requires test takers to sample any corresponding information from the correct overarching category of information (i.e., all rivers in Germany). However, both types of test requirements might be distinguished regarding their requirements of cognition (e.g., convergent vs. divergent cognition, Guilford, 1967; Nusbaum & Silvia, 2011). Both items require test takers not only to have the relevant information stored readily in long-term or semantic memory (i.e., crystallized intelligence), but they also require the test takers to be able to retrieve these information on command (i.e., retrieval ability). Arguably, both items hence tap the same underlying crystallized information structure in long-term or semantic memory; namely, the part where the information about rivers in Germany are stored. This example thus describes the immanent kinship of crystallized intelligence and broad retrieval ability and demonstrates that both constructs can be considered to stand in a reciprocal relationship with one another. Performing well on a Gc test (e.g., a declarative knowledge test) requires being able to retrieve the correct information from long-term

memory (i.e., Gr), and performing well on a Gr test (e.g., a verbal fluency test) requires to have stored (semantic) information in long-term memory (i.e., Gc).

In what ways are the results presented in manuscript IV significant for advancing our understanding regarding declarative knowledge and, thus, possibly the Gc factor itself? As outlined above, the cognitive requirements of both prototypical verbal fluency and prototypical declarative knowledge items are strikingly similar, and as such, the prerequisites of Gc and Gr should be akin: driving mechanisms of individual differences in Gr should also, at least to a considerable amount, drive individual differences in Gc. Previous literature has repeatedly questioned the very interpretation of the Gc factor, as it is proposed in recent intelligence models (e.g., CHC model, W. J. Schneider & McGrew, 2018; but c.f., Wilhelm & Kyllonen, 2021). For example, it was debated whether Gc should be interpreted as reflecting real individual differences in declarative knowledge or whether Gc is more reflective of differences in the investment traits and context variables of individuals' lives that aid knowledge acquisition over the course of a lifespan (c.f., Schipolowski, 2014). Through the framework presented in this dissertation, which puts Gc in the context of Gr, and through the empirical results presented in manuscript IV (Goecke et al., 2023), another possibility that should be considered is that Gc is reflective of the driving mechanisms (i.e., cognitive processes) that help test takers to access their knowledge in test situations. For example, just as considerable amounts of variance in Gr are explained by mental speed, working memory capacity, secondary memory, and Gc, considerable amounts of variance in Gc should be explicable through the same variables and Gr. In fact, using the data presented in manuscript IV, we find that 58.1% of the variance is explained in a latent factor of Gc (indicated through both vocabulary and declarative knowledge) through latent factors of mental speed, working memory capacity, secondary memory, and Gr (*CFI* = .927; *RMSEA* = .044; *SRMR* = .049). Taken together, furthering our understanding of Gr and its driving mechanisms will also further our understanding of Gc.

Being Wise About Studying Knowledge and Related Constructs

Overclaiming About Overclaiming

Research on psychological constructs should transcend the measures we apply. The initial idea of the overclaiming technique (that is, implementing foils in self-reports of knowledge, c.f., Paulhus et al., 2003) was to find a remedy for biased estimates of self-reported knowledge, or in other words, a measurement for "self-enhancement independent of ability" (Paulhus et al., 2003, p. 890). Although numerous studies have applied overclaiming questionnaires and studied overclaiming as understood by signal detection analytical approaches in numerous contexts (low- and high-stakes settings), the actual construct of overclaiming (i.e., individuals claim to know something they cannot or do not know) is still not yet well understood and open questions remain; for example, questions about the hitherto disappointing convergent validity of overclaiming with other constructs of the personality and cognitive ability framework. After all, the lack of convergent validity can certainly not be attributed to a lack of individual differences in overclaiming behavior as measured by OCQs.

However, I would like to point out that although some progress in understanding overclaiming as a construct of individual differences has been made (e.g., Ackerman & Ellingsen, 2014; Atir et al., 2015; Bensch et al., 2017, 2019; Dunlop et al., 2017, 2020; Grosz et al., 2017; Hülür et al., 2011, 2014; Müller & Moshagen, 2018; Vonkova et al., 2018; Ziegler et al., 2013; etc.), previous studies—including the studies presented in this dissertation—should have gone beyond a mono-methodological application (Donaldson & Grant-Vallone, 2002) of OCQs for assessing overclaiming (for exceptions, see e.g., Ackerman & Ellingsen, 2014; Raubenheimer, 1925). Overclaiming, in theory, should go beyond what is measured with overclaiming questionnaires. As discussed in the second manuscript interesting alternative measurement approaches would interconnect ability assessment of declarative knowledge and assessments of overclaiming. Thus, multi-method approached (Campbell & Fiske, 1959) studies are important if we want to obtain a more holistic picture of the construct of overclaiming rather than a measurement that is equated with the construct in question.

OCQs are one (and arguably the most prominent) way to measure overclaiming, but they might not be the only ones (c.f., manuscript II, Goecke et al., 2022). Equating overarching constructs (such as the propensity to overclaim) with individual measurement instruments (such as the OCQ-150 by Paulhus et al., 2003, or any other overclaiming questionnaire) may not fully serve the scientific community's interests. This can be problematic, as it limits the generalizability of findings and the scope of the construct studied. In other words, it is beneficial for the field that results can be attributed to the underlying constructs rather than the variance produced by a specific measurement instrument. Nevertheless, the efforts to develop complementary or alternative measurement instruments are sparse. Therefore, studying the phenomenon of overclaiming should be based on multivariate, and especially multi-method study designs, as proposed in manuscript II of this dissertation (Goecke et al., 2022). Enhancing our understanding of overclaiming, something that can arguably be observed in human behavior in our everyday lives (think of that friend who told you they read that book; or think about Donald Trump), might be helpful for more applied contexts, like job applications (e.g., Dunlop et al., 2020) and might thus improve our understanding of knowledge.

In line with this research desideratum, a suitable statistical measurement model for overclaiming needs to be established. While I acknowledge that previous scoring approaches of OCQs based on signal detection theory might be appealing due to their basic idea stemming from research traditions more experimental in nature (c.f., Kellen & Klauer, 2018), studying relatively simple behaviors like overclaiming (e.g., rate familiarity with foil vs. do not rate familiarity with foil) should be based on the purest of possible measurement approaches; for example, the raw familiarity ratings of foils (c.f., manuscript I, Goecke et al., 2020). Recently, a third scoring approach based on item-response theory has been proposed in the literature (c.f., Jin et al., 2023), and a future study comparing the different scoring approaches based on systematically manipulated questionnaire characteristics (e.g., the proportion of foils to reals) might be informative as to how the different scoring approaches might have affected previously reported results. In addition to that, further methodological

advances regarding the separation of effects of deliberate self-enhancement and mere positivity biases (i.e., condition-based regression analysis; e.g., Humberg et al., 2018) might be helpful in examining how overclaiming behavior might be explained.

The Idiosyncratic Structure of Knowledge and Semantic Memory

According to the previous literature, knowledge acquisition is an idiosyncratic process (e.g., Ackerman, 1996; Cattell, 1987; von Stumm, 2013; Wilhelm & Schroeders, 2019; Ziegler et al., 2012). Although formal education curricula provide relatively similar learning opportunities for some persons of specific regions in the world at the beginning of their lives (c.f., Watrin et al., 2023), it is easily conceivable that these same learning opportunities diversify once formal education is finished (e.g., Cattell, 1971; Rolfhus & Ackerman, 1996; Wilhelm & Schroeders, 2019). One consequence of this diversification of learning opportunities (e.g., pursuing a specific vocational training or visiting an art museum), mixed with individual differences in investment traits (e.g., von Stumm & Ackerman, 2013), is that in addition to general knowledge accumulation, individuals accumulate more domain-specific knowledge over time (e.g., Rusche & Ziegler, 2022). The accumulation of domain-specific knowledge should also be visible in continuously changing semantic memory structures (e.g., Bieth et al., 2021; Kenett & Thompson-Schill, 2020; Wulff et al., 2019; Yee & Thompson-Schill, 2016), because knowledge can be assumed to be stored semantically (Collins & Loftus, 1975; Smith et al., 1974; Tulving, 1972).

It makes sense to assume that persons being exposed to specific life events (e.g., Achaa-Amankwaa et al., 2022) or daily realities (i.e., jobs; Beier & Ackerman, 2003) have more or less knowledge regarding certain topics, for example, health knowledge. This idea is not new. Usually, participants are tested regarding their ability to either take advantage of their knowledge (i.e., respond correctly to a knowledge question) or not. With these kinds of studies, we are able to infer how declarative knowledge is structured cross-sectionally (e.g., Steger et al., 2019), but we learn only little about the idiosyncratic structure of knowledge in our subjects, although this poses an interesting field of research. Take, for example, the twins Person A and Person B. Both individuals might have attended the same school (i.e., similar

learning opportunities), but in addition to them having different interests (Rolfhus & Ackerman, 1996), they decided to study different subjects at different universities after successfully finishing the school system. Whereas Person A decided to study psychology, person B decided to study biology. Thus, their learning environments differed after that decision, and due to that the knowledge they will accumulate in line of their studies will differ. Person A will be an expert in psychology, but might remain a novice in biology, whereas the reverse is true for Person B. The consequence of this for conventional declarative knowledge tests is easily explained: whereas both persons might not differ much regarding knowledge domains tapping more general knowledge like basic geography knowledge, Person A will likely perform better on a test comprised of specific psychology questions, whereas person B will likely perform better on a test comprised of specific biology questions. The question why Person A outperforms Person B in a psychology-related knowledge test seems obvious on the surface; however, it might be more complicated.

One possibility for answering this question comes from the field of cognitive network science or semantic network structure research (e.g., Siew et al., 2019), because declarative knowledge might be viewed as a form of semantic network that is stored in semantic memory (Collins & Loftus, 1975; Smith et al., 1974; Tulving, 1972). Apart from the mere fact that Person A was exposed more to psychology-related content, the response to the question of why Person A outperforms Person B might be found in how Person A has structured their psychology-related knowledge in semantic memory, as compared to Person B. This idea has been frequently proposed in the literature regarding studies of expertise (Persky & Robinson, 2017; Wolff et al., 2017) and has recently gained tailwind through methodological advances in estimating semantic network structures (for an overview, see Christensen & Kenett, 2021 and Siew et al., 2019) depicting (domain-specific) knowledge (Siew & Guru, 2022). These semantic network approaches have long since made their way into the study of creativity (c.f. Benedek et al., 2017, 2023) and should not be mistaken for psychometric network approaches where covariances of single items or test scores are modeled (e.g., Epskamp et al., 2017). Taken together, this line of research offers an interesting opportunity to further our understanding

of Gc by investigating immanent differences between the knowledge representations on both the group and the individual level, which might be in turn connected with other sample characteristics like age (Cosgrove et al., 2023), health-related variables (Cosgrove et al., 2021), or individual characteristics, such as interests and exposure to learning opportunities, e.g. in the context of ecological momentary assessments (e.g., Shiffman et al., 2008).

How could this work? In order to gain insights into the "network structure of domain-specific knowledge" (Siew & Guru, 2022, p.1), Siew and Guru (2022) proposed a combination of behavioral and computational methods to unravel the semantic knowledge structure on the group level. In their study, the authors combined classical verbal fluency tasks (e.g., "Name as many psychology-related words as you can") with semantic network estimation methods (c.f., Christensen & Kenett, 2021) in order to compare groups of experts (e.g., psychology students) and novices (high-school students) regarding their semantic network structure. In line with the general argument of this dissertation, the authors assumed that verbal fluency tasks might offer sufficient approximations of knowledge regarding certain topics because these tasks tap the same underlying crystallized information structure as more conventional declarative knowledge tasks. Further, the authors hypothesized that differences in semantic network parameters (e.g., shortest average path lengths, clustering coefficients, and modularity indices) should be indicative of different knowledge levels (i.e., experts vs. novices). With that, the authors built upon the previous literature and offered a new way of quantifying knowledge and its semantic representation in memory on the group level. The only obvious drawback of such group comparisons is that the so far used methodology only allows for dichotomized considerations of groups (i.e., experts vs. novices). However, it might be more interesting to consider the more continuous nature of moderator variables such as knowledge in order to prevent a priori mis-categorizations and loss of information (MacCallum et al., 2002). In other words, future developments of semantic network estimation should consider the implementation of non-categorical moderator variables, for example, akin to how continuous moderator variables are already used in structural equation modeling (c.f., Local Structural Equation Modeling, Hildebrandt et al., 2016). These efforts

have only recently been extended to semantic network estimation on the individual level (e.g., Cosgrove et al., 2023), offering even a new path for examining individual differences in knowledge.

I argue that these advances could be understood as complementary to conventional knowledge tests of maximal ability because they offer an increment in information. Estimating semantic network structures regarding certain domains (e.g., psychology) might be helpful in understanding why certain persons outperform others in tests of declarative knowledge apart from previous arguments of more or less exposition to learning opportunities (Schroeders et al., 2015). In other words, gaining insight into the idiosyncratic knowledge structure of individuals by using semantic network estimation methods offers a helpful quantification for understanding why the mere application of knowledge (i.e., responding to knowledge questions) might be more or less hard sometimes. A common assumption, for example, is that experts on certain topics have "more sophisticated cognitive structures" regarding their knowledge (Siew & Guru, 2020, p. 2). How far these network structures might be differentially related to other constructs of individual differences (e.g., working memory capacity) remains an open question and offers an interesting research endeavor for the future.

Conclusion

In the vein of the dissertation at hand, I have presented four manuscripts that outlined a) how self-reported knowledge should be viewed in the light of existing ability test instruments, b) how response formats affect measures of declarative knowledge, and c) why broad retrieval ability and declarative knowledge should be understood as two edges of the same sword. With that, my work provides meaningful insights into the study of individual differences in knowledge and articulates the urge to keep investigating individual differences in knowledge from a broad perspective for the sake of an integral understanding of the same.

References

Achaa-Amankwaa, P., Steger, D., Wilhelm, O., & Schroeders, U. (2022). *Public Events Knowledge in an Age-Heterogeneous Sample: Reminiscence Bump or Bummer?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/yg3mf

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence*, *22*(2), 227–257. https://doi.org/10.1016/S0160-2896(96)90016-1

Ackerman, P. L., & Ellingsen, V. J. (2014). Vocabulary overclaiming — A complete approach: Ability, personality, self-concept correlates, and gender differences. *Intelligence*, *46*, 216–227. https://doi.org/10.1016/j.intell.2014.07.003

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, *26*(8), 1295–1303. https://doi.org/10.1177/0956797615588195

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*(2), 439–448. https://doi.org/10.1037/0022-3514.84.2.439

Benedek, M., Beaty, R. E., Schacter, D. L., & Kenett, Y. N. (2023). The role of memory in creative ideation. *Nature Reviews Psychology*, *2*. https://doi.org/10.1038/s44159-023-00158-z

Benedek, M., Kenett, Y. N., Umdasch, K., Anaki, D., Faust, M., & Neubauer, A. C. (2017). How semantic memory structure and intelligence contribute to creative thought: A

network science approach. *Thinking & Reasoning*, *23*(2), 158–183. https://doi.org/10.1080/13546783.2016.1278034

Bensch, D., Maaß, U., Greiff, S., Horstmann, K. T., & Ziegler, M. (2019). The nature of faking: A homogeneous and predictable construct? *Psychological Assessment*, *31*(4), 532–544. https://doi.org/10.1037/pas0000619

Bensch, D., Paulhus, D. L., Stankov, L., & Ziegler, M. (2017). Teasing Apart Overclaiming, Overconfidence, and Socially Desirable Responding. *Assessment*, *26*, 351–363. https://doi.org/10.1177/1073191117700268

Bieth, T., Kenett, Y., Ovando-Tellez, M., Lopez-Persem, A., Lacaux, C., Oudiette, D., & Volle, E. (2021). *Dynamic changes in semantic memory structure support successful problem-solving* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/38b4w

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

Butler, J. R., & Nelson, N. L. (2021). Children overclaim more knowledge than adults do, but for different reasons. *Journal of Experimental Child Psychology*, *201*, 104969. https://doi.org/10.1016/j.jecp.2020.104969

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cattell, R. B. (1958). What is "objective" in "objective personality tests?" *Journal of Counseling Psychology*, *5*(4), 285–289. https://doi.org/10.1037/h0046268

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Houghton Mifflin Company.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. Elsevier Science Publishers B.V.

Christensen, A. P., & Kenett, Y. N. (2021). Semantic network analysis (SemNA): A tutorial on preprocessing, estimating, and analyzing semantic networks. *Psychological Methods*. https://doi.org/10.1037/met0000463

Collins, A. M., & Loftus, E. F. (1975). A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Cosgrove, A. L., Beaty, R. E., Diaz, M. T., & Kenett, Y. N. (2023). Age differences in semantic network structure: Acquiring knowledge shapes semantic memory. *Psychology and Aging*. https://doi.org/10.1037/pag0000721

Cosgrove, A. L., Kenett, Y. N., Beaty, R. E., & Diaz, M. T. (2021). Quantifying flexibility in thought: The resiliency of semantic networks differs across the lifespan. *Cognition*, *211*, 104631. https://doi.org/10.1016/j.cognition.2021.104631

Cotter, K. N., Rodriguez-Boerwinkle, R. M., Christensen, A. P., Fekete, A., Smith, J. K., Smith, L. F., Tinio, P. P. L., & Silvia, P. J. (2023). Updating the Aesthetic Fluency Scale: Revised long and short forms for research in the psychology of the arts. *PLOS ONE*, *18*(2), e0281547. https://doi.org/10.1371/journal.pone.0281547

Cronbach, L. J. (1949). *Essentials of Psychological Testing*. Harper & Brothers, Publishers.

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding Self-Report Bias in Organizational Behavior Research. *Journal of Business and Psychology*, *17*(2), 245–262. https://doi.org/10.1023/A:1019637632584

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

Dunlop, P. D., Bourdage, J. S., de Vries, R. E., McNeill, I. M., Jorritsma, K., Orchard, M., Austen, T., Baines, T., & Choe, W.-K. (2020). Liar! Liar! (when stakes are higher): Understanding how the overclaiming technique can be used to measure faking in personnel selection. *Journal of Applied Psychology*, *105*(8), 784–799. https://doi.org/10.1037/apl0000463

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized Network Psychometrics: Combining Network and Latent Variable Models. *Psychometrika*, *82*(4), 904–927. https://doi.org/10.1007/s11336-017-9557-x

Fell, C. B., König, C. J., Jung, S., Sorg, D., & Ziegler, M. (2017). Are country level prevalences of rule violations associated with knowledge overclaiming among students? *International Journal of Psychology*, *54*(1), 1–6. https://doi.org/10.1002/ijop.12441

Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, *138*(2), 296–321. https://doi.org/10.1037/a0026556

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

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence*, *80*, 1–9. https://doi.org/10.1016/j.intell.2020.101449

Goecke, B., Schroeders, U., Zettler, I., Schipolowski, S., Golle, J., & Wilhelm, O. (2022). The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children. *Journal of Personality Assessment*, 1–12. https://doi.org/10.1080/00223891.2022.2144332

Goecke, B., Staab, M., Schittenhelm, C., & Wilhelm, O. (2022). Stop Worrying about Multiple-Choice: Fact Knowledge Does Not Change with Response Format. *Journal of Intelligence*, *10*(4), 1–22. https://doi.org/10.3390/jintelligence10040102

Goecke, B., Weiss, S., Steger, D., Schroeders, U., & Wilhelm, O. (2020). Testing competing claims about overclaiming. *Intelligence*, *81*, 1–11. https://doi.org/10.1016/j.intell.2020.101470

Goecke, B., Weiss, S., & Wilhelm, O. (2023). *Driving Factors of Individual Differences in Broad Retrieval Ability: Gr Is More Than the Sum of Its Parts* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/zy7mp

Grosz, M. P., Lösch, T., & Back, M. D. (2017). The narcissism-overclaiming link revisited. *Journal of Research in Personality*, *70*, 134–138. https://doi.org/10.1016/j.jrp.2017.05.006

Gruenewald, P. J., & Lockhead, G. R. (1980). The Free Recall of Category Examples. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(3), 225–240. https://doi.org/10.1037/0278-7393.6.3.225

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill, Inc.

Herreen, D., & Zajac, I. (2017). The Reliability and Validity of a Self-Report Measure of Cognitive Abilities in Older Adults: More Personality than Cognitive Function. *Journal of Intelligence*, *6*(1), 1–15. https://doi.org/10.3390/jintelligence6010001

Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., & Wilhelm, O. (2016). Exploring Factor Model Parameters across Continuous Variables with Local Structural Equation Models. *Multivariate Behavioral Research*, *51*(2–3), 257–258. https://doi.org/10.1080/00273171.2016.1142856

Hofer, G., Mraulak, V., Grinschgl, S., & Neubauer, A. C. (2022). Less-Intelligent and Unaware? Accuracy and Dunning–Kruger Effects for Self-Estimates of Different

Aspects of Intelligence. *Journal of Intelligence*, *10*(1), 1–18. https://doi.org/10.3390/jintelligence10010010

Holland, J. L. (1977). *Self-Directed Search: A guide to educational and vocational planning.* Consulting Psychologists Press.

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

Hülür, G., Wilhelm, O., & Schipolowski, S. (2014). Corrigendum to "Prediction of self-reported knowledge with over-claiming, fluid and crystallized intelligence and typical intellectual engagement" [Learning and Individual Differences 21 (6) (2011) 742–746]. *Learning and Individual Differences*, *35*, 153–154. https://doi.org/10.1016/j.lindif.2014.07.014

Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., van Zalk, M. H. W., Denissen, J. J. A., Nestler, S., & Back, M. D. (2018). Enhanced versus simply positive: A new condition-based regression analysis to disentangle effects of self-enhancement from effects of positivity of self-view. *Journal of Personality and Social Psychology*, *114*(2), 303–322. https://doi.org/10.1037/pspp0000134

Institute for Behavioral Research in Creativity. (1968). *Alpha Biographical Inventory.* Prediction Press.

Jin, K.-Y., Paulhus, D. L., & Shih, C.-L. (2023). A New Approach to Desirable Responding: Multidimensional Item Response Model of Overclaiming Data. *Applied Psychological Measurement*, 014662162311517. https://doi.org/10.1177/01466216231151704

Kellen, D., & Klauer, K. C. (2018). Elementary Signal Detection and Threshold Theory. In J. T. Wixted (Ed.), *Steven' Handbook of Experimental Psychology and Cognitive Neuroscience* (Fourth Edition). John Wiley & Sons.

Kenett, Y., & Thompson-Schill, S. L. (2020). *Novel conceptual combination can dynamically reconfigure semantic memory networks* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/crp47

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of abilities: A review and meta-analysis. *Journal of Applied Psychology*, *67*(3), 280–296. https://doi.org/10.1037/0021-9010.67.3.280

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037//1082-989X.7.1.19

Müller, S., & Moshagen, M. (2018). Overclaiming shares processes with the hindsight bias. *Personality and Individual Differences*, *134*, 298–300. https://doi.org/10.1016/j.paid.2018.06.035

Nusbaum, E. C., & Silvia, P. J. (2011). Are intelligence and creativity really so different?☆Fluid intelligence, executive processes, and strategy use in divergent thinking. *Intelligence*, *39*(1), 36–45. https://doi.org/10.1016/j.intell.2010.11.002

OECD. (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD. https://doi.org/10.1787/9789264190511-en

Olderbak, S., & Wilhelm, O. (2020). Overarching Principles for the Organization of Socioemotional Constructs. *Current Directions in Psychological Science*, *29*(1), 63–70. https://doi.org/10.1177/0963721419884317

Paulhus, D. L. (2011). Overclaiming on personality questionnaires. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessement* (pp. 151–164). Oxford University Press, Inc.

Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence*, *32*(3), 297–314. https://doi.org/10.1016/j.intell.2004.02.001

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

Persky, A. M., & Robinson, J. D. (2017). Moving from Novice to Expertise and Its Implications for Instruction. *American Journal of Pharmaceutical Education*, *81*(9), 6065. https://doi.org/10.5688/ajpe6065

Phillips, D. L., & Clancy, K. J. (1972). Some effects of "social desirability" in survey studies. *American Journal of Sociology*, *77*(5), 921–940. https://doi.org/10.1086/225231

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Randall, D. M., & Fernandes, M. E. (1991). The social desirability response bias in ethics research. *Journal of Business Ethics*, *10*, 805–817. https://doi.org/10.1007/BF00383696

Raubenheimer, A. S. (1925). An experimental study of some behavior traits of the potentially delinquent boy. *Psychological Monographs*, *34*, 1–107.

Reiss, S., Tillmann, A., Schreiner, M., Schweizer, K., Krömker, D., & Moosbrugger, H. (2009). Online-Self-Assessments zur Erfassung studienrelevanter Kompetenzen. *Zeitschrift für Hochschulentwicklung*, *4*(1), 60–71. https://doi.org/10.3217/ZFHE-4-01/05

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. https://doi.org/10.1037/0022-0663.88.1.174

Rosen, V. M., & Engle, R. W. (1997). The Role of Working Memory Capacity in Retrieval. *Journal of Experimental Psychology: General*, *126*(3), 211–227. https://doi.org/10.1037/0096-3445.126.3.211

Rusche, M. M., & Ziegler, M. (2022). The interplay between domain-specific knowledge and selected investment traits across the life span. *Intelligence*, *92*, 101647. https://doi.org/10.1016/j.intell.2022.101647

Schneider, W. (2008). The Development of Metacognitive Knowledge in Children and Adolescents: Major Trends and Implications for Education. *Mind, Brain, and Education*, *2*(3), 114–121. https://doi.org/10.1111/j.1751-228X.2008.00041.x

Schneider, W. J., & McGrew, K. S. (2018). The Cattell–Horn–Carroll Theory of Cognitive Abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 73–163). The Guilford Press.

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., Schipolowski, S., & Wilhelm, O. (2020). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 5. Bis 7. Klasse (BEFKI 5-7)*. Hogrefe Verlag.

Schroeders, U., Schipolowski, S., Zettler, I., Golle, J., & Wilhelm, O. (2016). Do the smart get smarter? Development of fluid and crystallized intelligence in 3rd grade. *Intelligence*, *59*, 84–95. https://doi.org/10.1016/j.intell.2016.08.003

Schroeders, U., Watrin, L., & Wilhelm, O. (2021). Age-related nuances in knowledge assessment. *Intelligence*, *85*, 101526. https://doi.org/10.1016/j.intell.2021.101526

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

Siew, C. S. Q., & Guru, A. (2022). Investigating the network structure of domain-specific knowledge using the semantic fluency task. *Memory & Cognition*. https://doi.org/10.3758/s13421-022-01314-1

Siew, C. S. Q., Wulff, D. U., Beckage, N. M., Kenett, Y. N., & Meštrović, A. (2019). Cognitive Network Science: A Review of Research on Cognition through the Lens of Network Representations, Processes, and Dynamics. *Complexity*, *2019*, 1–24. https://doi.org/10.1155/2019/2108423

Smith, E. E., Rips, L. J., & Shoben, E. J. (1974). Semantic Memory and Psychological Semantics. *Psychology of Learning and Motivation*, *8*, 1–45. https://doi.org/10.1016/S0079-7421(08)60451-X

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72–99. https://doi.org/10.2307/1412159

Steger, D., Jankowsky, K., Schroeders, U., & Wilhelm, O. (2022). The Road to Hell Is Paved With Good Intentions: How Common Practices in Scale Construction Hurt Validity. *Assessment*, *00*(0), 1–14. https://doi.org/10.1177/10731911221124846

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

Steger, D., Schroeders, U., & Wilhelm, O. (2021). Caught in the act: Predicting cheating in unproctored knowledge assessment. *Assessment*, *28*(3), 1004–1017.

Tulving, E. (1972). Episodic and Semantic Memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory*. Academic Press.

Tulving, E., & Watkins, M. J. (1973). Continuity between Recall and Recognition. *The American Journal of Psychology*, *86*(4), 739–748. https://doi.org/10.2307/1422081

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2011). Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *The Quarterly Journal of Experimental Psychology*, *64*(3), 447–466. https://doi.org/10.1080/17470218.2010.505292

von Stumm, S. (2013). Investment Traits and Intelligence in Adulthood: Assessment and Associations. *Journal of Individual Differences*, *34*(2), 82–89. https://doi.org/10.1027/1614-0001/a000101

von Stumm, S., & Ackerman, P. L. (2013). Investment and intellect: A review and meta-analysis. *Psychological Bulletin*, *139*(4), 841–869. https://doi.org/10.1037/a0030746

Vonkova, H., Papajoanu, O., & Stipek, J. (2018). Enhancing the Cross-Cultural Comparability of Self-Reports Using the Overclaiming Technique: An Analysis of Accuracy and Exaggeration in 64 Cultures. *Journal of Cross-Cultural Psychology*, *49*(8), 1247–1268. https://doi.org/10.1177/0022022118787042

Watrin, L., Schroeders, U., & Wilhelm, O. (2023). Gc at its boundaries: A cross-national investigation of declarative knowledge. *Learning and Individual Differences*, *102*, 102267. https://doi.org/10.1016/j.lindif.2023.102267

Weiss, S., & Roberts, R. D. (2018). Using Anchoring Vignettes to Adjust Self-Reported Personality: A Comparison Between Countries. *Frontiers in Psychology*, *9*, 325. https://doi.org/10.3389/fpsyg.2018.00325

Weiss, S., Steger, D., Kaur, Y., Hildebrandt, A., Schroeders, U., & Wilhelm, O. (2020). On the Trail of Creativity: Dimensionality of Divergent Thinking and Its Relation With Cognitive Abilities, Personality, and Insight. *European Journal of Personality*, *35*, 22. https://doi.org/10.1002/per.2288

Wilhelm, O., & Kyllonen, P. (2021). To predict the future, consider the past: Revisiting Carroll (1993) as a guide to the future of intelligence research. *Intelligence*, *89*, 101585. https://doi.org/10.1016/j.intell.2021.101585

Wilhelm, O., & Schroeders, U. (2019). Intelligence. In R. Sternberg & J. Funke (Eds.), *The Psychology of Human Thought: An Introduction* (pp. 257–277). Heidelberg University Publishing.

Wolff, C. E., Jarodzka, H., & Boshuizen, H. P. A. (2017). See and tell: Differences between expert and novice teachers' interpretations of problematic classroom management events. *Teaching and Teacher Education*, *66*, 295–308. https://doi.org/10.1016/j.tate.2017.04.015

Wulff, D. U., De Deyne, S., Jones, M. N., & Mata, R. (2019). New Perspectives on the Aging Lexicon. *Trends in Cognitive Sciences*, *23*(8), 686–698. https://doi.org/10.1016/j.tics.2019.05.003

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, *23*(4), 1015–1027. https://doi.org/10.3758/s13423-015-0948-7

Zell, E., & Krizan, Z. (2014). Do People Have Insight Into Their Abilities? A Metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111–125. https://doi.org/10.1177/1745691613518075

Ziegler, M., Danay, E., Heene, M., Asendorpf, J., & Bühner, M. (2012). Openness, fluid intelligence, and crystallized intelligence: Toward an integrative model. *Journal of Research in Personality*, *46*(2), 173–183. https://doi.org/10.1016/j.jrp.2012.01.002

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The Vocabulary and Overclaiming Test (VOC-T). *Journal of Individual Differences*, *34*(1), 32–40. https://doi.org/10.1027/1614-0001/a000093

VII Appendix

Testing Competing Claims About Overclaiming

Supplementary Material

Descriptive Statistics OCQ

SM Table 1

Descriptive Statistics of All OCQ Items.

Real=1, Foil=2	Parcel	Item	*M*	*SD*	Skew	Kurtosis	*M* (Rating)	*SD* (Rating)	*M (SD)* (Hülür et al., 2011)
1	1	Le Corbusier	1.53	.87	1.72	2.49	12.20	3.60	
2	1	Opulentismus	1.56	.81	1.65	2.92	3.83	.90	
1	2	Akropolis	2.89	1.18	.07	-.86	6.33	.75	
1	3	Bauhaus	3.35	1.12	-.07	-.93	9.83	2.19	
1	4	Kork	3.14	1.32	-.13	-1.15	5.00	.89	
1	5	Bitumen	1.76	1.13	1.52	1.41	14.60	3.07	
1	1	Zöliakie	2.47	1.52	.52	-1.24	12.40	2.94	
1	2	Ernährungs-pyramide	4.11	.95	-.99	.64	4.83	1.57	
1	3	Frutarier	3.81	1.27	-.84	-.35	7.50	.76	
1	4	Veganismus	4.48	.80	-1.52	1.75	4.67	1.25	
1	5	Spurenelemente	3.26	1.13	.11	-.96	8.17	2.27	
2	1	Ökomanie	2.07	1.18	1.08	.31	3.50	.50	
1	1	Stummfilm	4.32	.86	-1.20	.99	4.80	1.17	
1	2	Charlie Chaplin	3.78	.97	-.53	-.04	7.80	.75	
2	1	Jack Matchman	1.26	.66	3.13	10.89	3.83	.90	
1	3	Quentin Tarantino	3.56	1.20	-.39	-.88	9.80	1.33	
1	4	4-D-Film	3.51	1.26	-.37	-.98	6.00	2.28	
1	1	Etagere	3.23	1.69	-.20	-1.67	11.60	4.22	
1	2	Kurkuma	3.79	1.35	-.80	-.62	10.20	1.60	
1	3	Essigreiniger	4.37	.85	-1.43	2.12	6.80	1.72	
2	2	Messerschleife	2.74	1.57	.30	-1.46	3.83	1.07	
1	4	Teflonbeschichtung	3.81	1.16	-.71	-.37	7.80	1.60	
1	5	Thermomix	4.15	.99	-.97	.14	8.00	1.79	
1	1	Bill Gates	3.78	.95	-.29	-.69	8.20	1.94	
1	2	IP-Adresse	3.35	1.04	-.07	-.48	8.40	1.20	
1	3	kompilieren	1.89	1.19	1.32	.77	12.60	2.50	
1	4	DNS Server	1.94	1.08	1.36	1.39	12.40	3.14	
2	1	Primulator	1.19	.50	3.28	14.13	3.67	.47	
1	5	Phosphoreszierende Garne	1.87	1.20	1.29	.60	15.20	1.94	
2	2	Balconette Schleife	1.85	1.12	1.18	.40	3.83	.69	
1	1	Androgyn	3.05	1.53	-.11	-1.47	11.40	1.85	

1	2	Jersey Stoffe	2.93	1.39	.08	-1.25	9.40	1.74	
1	3	Vivienne Westwood	2.21	1.29	.85	-.39	12.60	1.62	
1	4	Katy Perry	3.98	.96	-.44	-.95	5.40	.49	
1	5	DJ Bobo	3.49	1.12	-.25	-.86	8.20	3.71	
1	1	Love Generation	2.12	1.16	.89	-.15	11.00	5.02	
2	3	Iradicals	1.43	.86	2.47	6.20	3.00	1.15	
1	2	Ed Sheeran	3.99	1.10	-.91	.01	5.80	1.33	
1	3	Pinterest	3.39	1.44	-.36	-1.25	8.00	1.10	
1	4	Sidestream	1.49	.91	2.21	4.75	12.80	2.64	
1	5	Twerken	3.39	1.49	-.42	-1.23	8.20	1.94	
2	4	Joiling	1.24	.67	3.52	13.44	3.67	.75	
1	2	Verafake	2.15	1.68	.92	-1.00	10.80	2.04	
2	5	Click Chocking	1.25	.75	3.69	14.12	3.83	.90	
1	2	Planking	2.91	1.64	.05	-1.64	8.40	1.36	
2	2	Ram-Rom-Wandler	1.29	.75	2.92	8.54	3.33	1.49	
1	2	Man-in-the-Middle-Angriff	1.36	.94	2.86	7.47	13.83	1.21	
1	3	Phishing	2.33	1.34	.61	-.91	10.33	1.70	1.43 (.85)
1	4	Photovoltaik	3.13	1.34	-.21	-1.04	10.67	1.80	1.34 (.83)
1	5	HDMI	3.46	1.04	-.03	-.84	9.33	2.21	
1	1	Prosa	3.35	1.19	-.12	-.92	8.00	.00	2.18 (1.55)
1	2	Kurt Tucholsky	2.40	1.16	.55	-.55	9.50	1.12	
1	3	Sokratischer Dialog	2.45	1.30	.64	-.70	12.83	1.77	
1	4	Ringparabel	2.71	1.32	.32	-1.11	9.83	1.34	
2	2	Epokritizismus	1.34	.69	2.65	8.49	4.17	.90	1.27 (.77)
1	5	Assonanz	1.68	.96	1.52	1.89	12.50	2.50	
1	1	Flachdruck	2.03	1.07	1.14	.86	9.83	1.86	
2	3	Flush-Technik	1.41	.75	2.28	6.17	4.33	.47	
2	4	Erwin Modur	1.24	.68	3.40	12.41	4.17	.69	
1	2	Komplementärfarbe	3.88	1.16	-.87	-.06	8.00	1.15	2.73 (1.58)
1	3	Pop Art	3.26	1.26	-.12	-1.07	10.00	.58	3.09 (1.45)
1	4	Vedutenmalerei	1.32	.69	2.70	8.61	14.00	2.65	
1	5	Claude Monet	3.03	1.16	.14	-.78	9.33	1.80	
1	1	Sonatenhauptsatzform	1.92	1.13	1.26	.82	12.17	2.41	
2	5	Laworette	1.23	.58	3.10	11.38	3.33	.94	1.45 (1.03)
1	2	Ouvertüre	2.64	1.14	.49	-.57	9.00	1.73	2.56 (1.42)
1	3	Mezzoforte	2.84	1.44	.22	-1.33	10.17	2.79	
1	4	Oktave	3.62	1.27	-.38	-1.12	6.50	2.06	3.81 (1.30)
1	5	Bratsche	3.37	1.26	-.20	-1.08	6.83	1.46	
1	1	Stammzelldebatte	2.92	1.34	.03	-1.15	9.67	.94	
1	2	Ideenlehre	2.55	1.26	.54	-.73	11.83	1.77	
1	3	kategorischer Imperativ	3.15	1.38	-.06	-1.29	11.83	2.19	2.43 (1.42)
2	1	Euphremetik	1.56	.82	1.90	4.31	4.33	.94	
1	4	Leib-Seele-Dualismus	2.69	1.45	.34	-1.26	13.50	3.04	
1	5	Epistemologie	2.04	1.18	1.10	.32	14.50	1.61	

1	1	Phänotyp	3.09	1.49	-.06	-1.43	10.17	1.67	1.41 (.89)
2	3	Syllabozin	1.22	.55	3.23	13.24	4.67	.47	
1	2	Segelklappen	1.76	1.12	1.55	1.60	12.33	2.92	
1	3	Leitbündel	1.80	1.13	1.47	1.36	14.50	2.93	
2	3	Nucleinkomplex	2.05	1.02	1.10	.89	4.83	.37	
1	1	Aromatische Kohlenwasserstoffe	2.04	1.10	1.12	.62	12.00	1.63	
1	2	Michael Faraday	2.06	1.22	.86	-.33	10.17	2.41	
1	3	Redox-Reaktion	2.57	1.27	.39	-.88	9.50	1.89	3.00 (1.60)
1	4	Anthroposophie	2.20	1.26	.75	-.56	16.33	2.56	
1	5	Aufenthaltsbestimmungsrecht	3.04	1.22	.08	-.94	9.50	.50	
2	4	Ralph Medward	1.22	.60	3.42	13.59	4.17	.37	
1	1	Autoritativer Erziehungsstil	3.49	1.27	-.46	-.79	11.00	1.83	
2	1	Odd Pot	1.13	.39	3.51	14.63	4.50	.76	
1	1	Bullenmarkt	1.55	.96	1.93	3.17	11.83	.90	
1	2	Diskontsatz	1.39	.75	2.33	6.08	14.33	1.37	
1	3	Kapitalertragsteuer	2.32	.97	.64	.13	11.83	1.57	2.19 (1.20)
1	4	Coriolis-Kraft	2.07	1.25	.92	-.35	10.33	1.37	
1	5	Sahelzone	2.06	1.26	.92	-.34	9.00	1.15	2.11 (1.50)
2	2	Scheibenblitz	1.43	.78	2.39	6.68	4.17	.69	1.80 (1.13)
1	1	Litosphäre	1.93	.96	1.09	1.01	10.67	1.70	
1	2	Wormser Konkordat	1.88	.99	1.34	1.69	11.17	1.77	
2	3	Elsässer Verträge	2.31	.94	.99	1.00	4.50	.50	
1	3	Vierte französische Republik	2.00	1.04	1.01	.48	13.17	1.77	
1	4	Vertrag von Maastricht	2.73	1.07	.62	-.48	12.50	1.98	
1	2	Approbation	2.91	1.43	.10	-1.31	14.00	1.26	
1	3	Zeckenkarte	3.14	1.46	-.13	-1.35	9.00	1.41	
2	5	Palliativmedizin	3.38	1.39	-.37	-1.14	11.60	3.01	
1	4	Erregerpräferenz	2.23	1.13	.86	.08	4.17	.37	
1	1	Phonetik	2.78	1.28	.33	-.98	10.60	2.06	
2	2	Satzstigma	1.67	.92	1.65	2.82	3.67	.47	
1	2	Diachronie	1.52	.75	1.56	2.55	12.20	3.37	
1	4	Blaise Pascal	1.77	.99	1.42	1.62	14.00	3.16	
1	5	Asymptote	3.12	1.35	-.08	-1.21	9.40	.80	
2	4	Annäherungs-komplexität	1.98	1.04	1.15	1.01	4.33	.47	
1	1	Vektor	3.43	1.22	-.26	-.95	9.40	1.62	
2	1	Zytonose	1.49	.96	2.28	4.82	4.50	.76	
1	5	Autoimmunerkrankung	3.50	1.20	-.30	-.83	9.50	1.61	1.74 (1.12)
1	1	Fraktur	3.80	1.12	-.66	-.39	10.33	1.80	2.57 (1.39)
1	2	Pneumonie	2.77	1.38	.36	-1.13	12.33	1.97	
1	3	Osmanisches Reich	2.86	1.05	.47	-.48	10.00	2.28	
1	4	Sansculotten	1.78	1.03	1.38	1.27	12.40	4.32	

1	5	Basken	2.37	1.01	.77	.21	11.00	2.28	
1	2	Diode	2.31	1.06	.77	.14	9.50	1.71	
2	5	Amaltonverfahren	1.28	.62	3.07	12.43	4.50	.50	
1	3	Glühemission	1.65	.95	1.75	2.92	11.83	2.03	
1	4	Winkelgeschwindigkeit	1.96	1.10	1.31	1.19	10.50	1.26	
1	5	Säkularismus	2.55	1.35	.49	-1.01	12.50	1.50	
1	1	Überhangmandat	2.83	1.25	.41	-.87	11.50	1.26	2.24 (1.30)
1	2	Legislaturperiode	3.58	1.26	-.52	-.79	9.67	1.11	1.85 (1.27)
1	3	Gabriele Krone-Schmalz	1.18	.64	4.20	18.88	14.00	2.90	
1	4	Caro Daur	1.57	1.25	2.02	2.50	11.00	3.16	
1	5	Stephen Hawking	3.79	1.03	-.56	-.22	9.20	1.94	
1	2	Introspektion	2.64	1.62	.40	-1.47	15.40	1.36	
1	3	Behaviorismus	3.39	1.50	-.40	-1.26	12.80	1.72	
1	4	systemische Therapie	2.78	1.40	.30	-1.18	13.60	1.62	
1	3	Genfer Konvention	2.47	1.10	.74	-.16	11.00	.82	1.81 (1.00)
1	4	Fernabsatzgesetz	1.68	.96	1.69	2.80	13.67	2.13	
1	5	Petitionsrecht	2.51	1.04	.64	-.09	12.50	1.71	
1	1	Kaaba	2.12	1.39	.97	-.43	10.50	1.50	1.96 (1.56)
2	3	Umberianer	1.21	.51	2.97	12.22	4.17	.69	
1	2	Trinitatiszeit	1.51	.90	2.11	4.38	12.33	2.81	
1	1	nordische Kombination	2.11	1.26	.92	-.26	11.00	1.67	
1	2	Caddy	2.15	1.28	.88	-.35	10.00	2.10	
1	3	Shotokan Karate	1.49	.90	2.21	4.91	9.60	1.96	
1	5	Effektstärke	2.95	1.45	.04	-1.35	14.80	2.32	
2	1	trigonometrische Verteilung	2.40	1.10	.76	-.11	4.33	.75	
1	1	Standardabweichung	3.73	1.22	-.61	-.64	11.80	2.99	
1	2	Korrelation	3.83	1.24	-.78	-.49	13.40	1.02	
1	1	Schadstoffregister	2.47	.94	.78	.51	NA	NA	
1	2	Umweltdruck	2.12	.93	.82	.81	11.60	3.01	
1	3	FCKW	2.49	1.31	.37	-1.06	11.40	4.72	
1	1	Tarifautonomie	2.12	1.08	.87	.09	12.67	1.11	
1	2	Handelsregister	2.45	1.04	.69	.03	12.17	2.11	2.49 (1.29)
2	4	Rotationsscheine	1.43	.64	1.66	3.73	4.50	.50	1.89 (1.09)

Note. All items shaded in grey were originally used in the study of Hülür and colleagues (2011).

Signal Detection Analysis

Following previous scoring procedures of overclaiming questionnaires (e.g., Paulhus et al., 2003; Paulhus & Harms, 2004; Paulhus & Petrusic, 2010), we dichotomized participants' responses not only at the lowest cut-off of the response-scale (i.e., 1 = unfamiliar; > 1 = familiar"), but across all cut-offs in order to stabilize the statistical indices (i.e., 1 & 2 = unfamiliar; > 2 = familiar" etc.). Four OC accuracy scores and four OC bias scores were computed and averaged to obtain signal detection parameter estimates (please see SM Figure 1 for a full correlation plot of the indices). For each cut-off, we calculated the *Hit-Ratio* (HR) and the *False-Alarm-Ratio* (FAR) as follows after dichotomizing:

$$HR = \frac{Hits + 0.5}{Signal\ Trials + 1} \qquad (1)$$

$$FAR = \frac{False\ Alarms + 0.5}{Noise\ Trials + 1} \qquad (2)$$

The formulas of the ratios were adjusted to ensure that indices are corrected for extreme hit-rates and extreme false-alarm-rates (i.e., 0 or 1; see Hautus, 1995). The following two indices were calculated within the framework of signal detection analysis. The *accuracy index d'* (d prime) and the response bias *index c* (criterion location, see also Macmillan, 1993; Stanislaw & Todorov, 1999):

$$d' = \Phi^{-1}(HR) - \Phi^{-1}(FAR) \qquad (3)$$

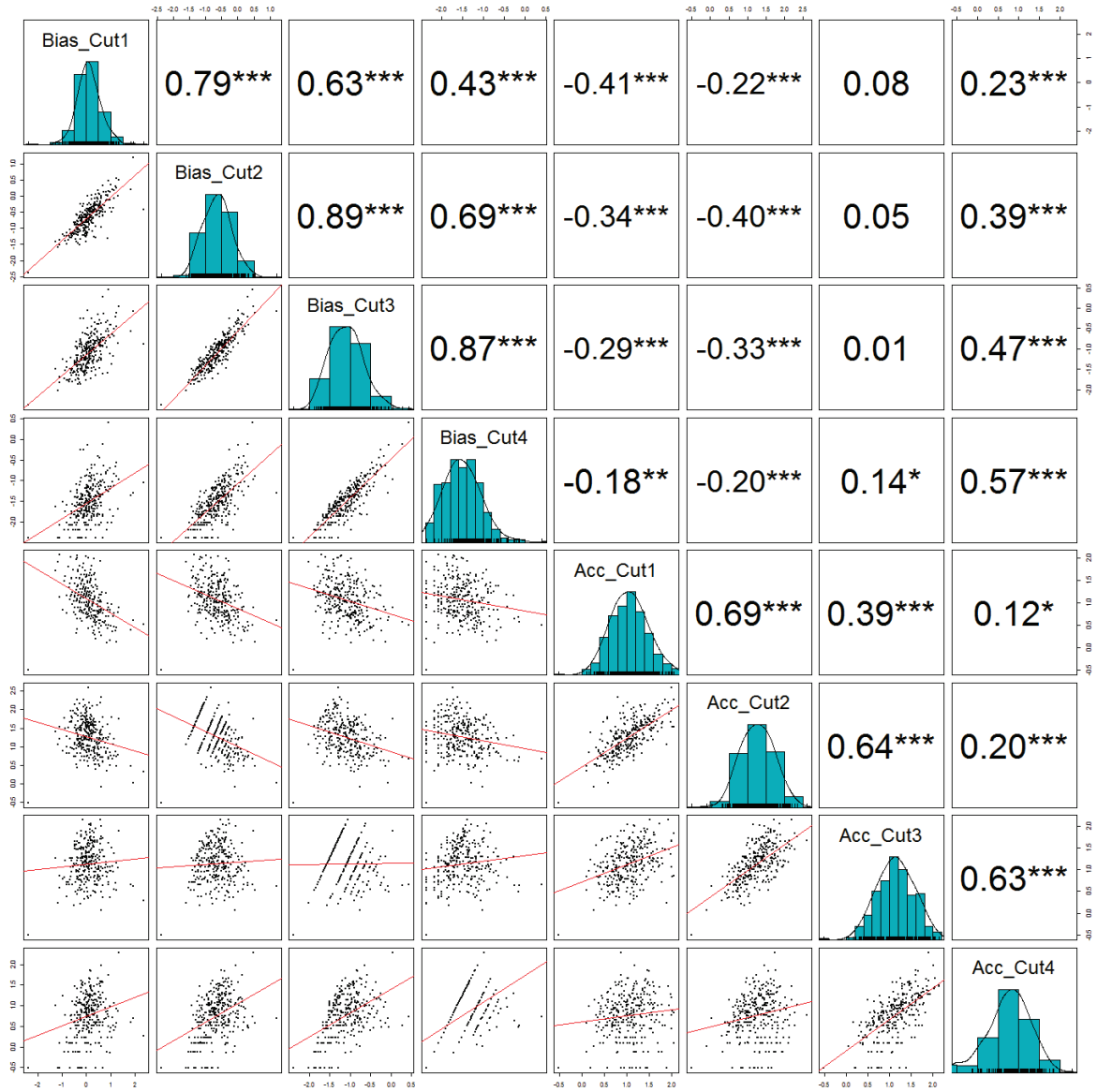
$$c = \frac{\Phi^{-1}(HR) + \Phi^{-1}(FAR)}{2} \qquad (4)$$

Please note that the correlation between OC accuracy and OC bias was -.01 (see SM Table 2). Therefore, the magnitude of this particular correlation fits into the hitherto heterogeneous picture of associations between OC accuracy and OC bias found across the

literature (r = .21, Paulhus et al., 2003; r = -.01, Dunlop et al., 2017; r = .15, Dunlop et al., 2019; r = .26 and r = -.74, Ziegler et al., 2013).

SM Figure 1

Correlation Plot of OC Accuracy and OC Bias Scores For All Four Cut-Offs



Note. Distributions are depicted on the diagonal. . $p < .001$ = ***, $p < .01$ = **, $p < .05$ = *.

SM Table 2

Correlation matrix of signal detection indices (d', c, hit-rate, and false alarm rate) and aggregate variables of fluid intelligence (gf), divergent thinking (DT), honesty-humility (HH), openness creativity, faking ability, mean familiarity ratings of foils (OC), mean familiarity ratings of reals (self-reported knowledge), and crystallized intelligence (gc). Additionally, correlations between signal detection indices and individual parcels of OC, SRK, and gc are depicted.

	Variable	*M (SD)*	1	2	3	4	5	6	7	8	9	10	11	12	13
1	OC	1.54 (.41)	1												
2	ocq.parcel1	1.55 (.43)	.86	1											
3	ocq.parcel2	1.70 (.57)	.84	.67	1										
4	ocq.parcel3	1.60 (.47)	.84	.68	.63	1									
5	ocq.parcel4	1.40 (.44)	.82	.62	.59	.62	1								
6	ocq.parcel5	1.47 (.51)	.86	.71	.62	.66	.67	1							
7	SRK	2.72 (.50)	.66	.59	.61	.61	.44	.53	1						
8	srk.parcel1	2.74 (.56)	.60	.54	.55	.58	.39	.46	.95	1					
9	srk.parcel2	2.68 (.50)	.66	.59	.60	.62	.43	.53	.95	.90	1				
10	srk.parcel3	2.67 (.51)	.65	.56	.60	.61	.41	.54	.94	.87	.89	1			
11	srk.parcel4	2.62 (.49)	.66	.57	.61	.59	.48	.53	.93	.85	.86	.85	1		
12	srk.parcel5	2.88 (.56)	.57	.52	.54	.52	.39	.43	.94	.87	.86	.85	.86	1	
13	gc	.58 (.10)	-.04	-.07	.01	.02	-.09	-.05	.37	.40	.37	.38	.27	.31	1

14	gc.parcel1	.55 (.14)	-.04	-.03	.01	.02	-.11	-.06	.30	.33	.30	.31	.19	.25	.75
15	gc.parcel2	.56 (.14)	.00	-.03	.02	.07	-.08	-.01	.29	.31	.29	.30	.22	.25	.78
16	gc.parcel3	.60 (.14)	-.03	-.06	.01	.00	-.05	-.05	.26	.26	.25	.29	.21	.22	.73
17	gc.parcel4	.58 (.13)	.01	-.04	.05	.04	-.01	-.01	.32	.35	.32	.29	.26	.29	.73
18	gc.parcel5	.60 (.13)	-.09	-.10	-.06	-.07	-.08	-.07	.17	.22	.19	.20	.10	.12	.67
19	gf	.61 (.16)	-.05	-.07	-.03	-.02	-.06	-.06	.17	.17	.15	.14	.17	.16	.30
20	DT	4.37 (.89)	.13	.07	.19	.11	.11	.05	.34	.34	.32	.28	.32	.33	.39
21	Openness Creativity	3.41 (.84)	.20	.21	.19	.16	.11	.16	.28	.25	.28	.31	.22	.27	.26
22	Honesty-Humility	3.35 (.81)	-.08	-.07	-.01	-.04	-.13	-.09	.05	.05	.03	.05	.05	.04	.07
23	Faking Ability	.34 (.14)	-.01	.01	.02	.03	-.08	-.03	.17	.23	.16	.16	.13	.13	.24
24	OC accuracy (*d'*)	4.22 (1.38)	-.45	-.40	-.38	-.31	-.42	-.41	.31	.31	.27	.27	.25	.34	.49
25	Hit Rate	1.71 (.50)	.66	.59	.61	.62	.44	.53	1	.95	.96	.95	.93	.93	.38
26	False Alarm Rate	.60 (.39)	1	.88	.85	.85	.80	.84	.67	.61	.67	.65	.66	.58	-.03
27	OC bias (*c*)	-3.13 (1.63)	.88	.78	.80	.76	.65	.71	.91	.85	.88	.86	.87	.83	.17

Note. $p < .05$ for all correlations > .12.

SM Table 2 to be continued

	Variable	*M (SD)*	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	OC	1.54 (.41)														
2	ocq.parcel1	1.55 (.43)														
3	ocq.parcel2	1.70 (.57)														
4	ocq.parcel3	1.60 (.47)														
5	ocq.parcel4	1.40 (.44)														
6	ocq.parcel5	1.47 (.51)														
7	SRK	2.72 (.50)														
8	srk.parcel1	2.74 (.56)														
9	srk.parcel2	2.68 (.50)														
10	srk.parcel3	2.67 (.51)														
11	srk.parcel4	2.62 (.49)														
12	srk.parcel5	2.88 (.56)														
13	gc	.58 (.10)														
14	gc.parcel1	.55 (.14)	1													
15	gc.parcel2	.56 (.14)	.53	1												
16	gc.parcel3	.60 (.14)	.40	.43	1											
17	gc.parcel4	.58 (.13)	.46	.49	.42	1										
18	gc.parcel5	.60 (.13)	.35	.38	.40	.32	1									

19	gf	.61 (.16)	.19	.22	.23	.22	.24	1								
20	DT	4.37 (.89)	.27	.35	.27	.32	.22	.29	1							
21	Openness Creativity	3.41 (.84)	.23	.20	.09	.19	.23	.08	.12	1						
22	Honesty-Humility	3.35 (.81)	.04	.08	-.04	.08	.08	.12	-.02	.22	1					
23	Faking Ability	.34 (.14)	.21	.22	.13	.17	.13	.20	.24	.09	.07	1				
24	OC accuracy (*d'*)	4.22 (1.38)	.40	.36	.34	.38	.33	.25	.25	.11	.14	.24	1			
25	Hit Rate	1.71 (.50)	.31	.30	.27	.34	.18	.17	.35	.29	.05	.18	.34	1		
26	False Alarm Rate	.60 (.39)	-.02	.01	-.03	.02	-.09	-.05	.14	.21	-.07	-.01	-.41	.67	1	
27	OC bias (*c*)	-3.13 (1.63)	.13	.16	.11	.18	.04	.06	.28	.26	-.01	.08	-.01	.91	.88	1

Note. $p < .05$ for all correlations > .12.

SM Table 3

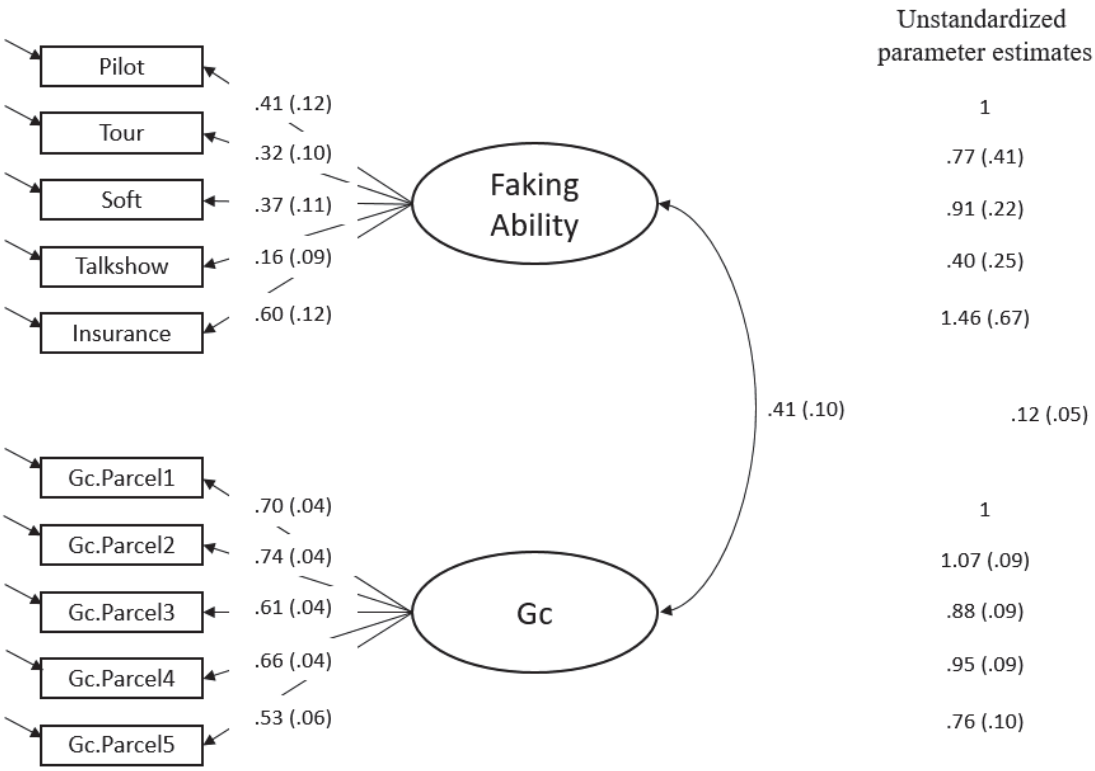
Regression Models With d' (OC Accuracy) and c (OC bias) As Predictors.

Criterion	Intercept			OC Accuracy (d')			OC Bias (c)					
	B	SE_B	p_B	β_{Acc}	SE_{Acc}	p_{Acc}	β_{Bias}	SE_{Bias}	p_{Bias}	R^2	R^2_{Acc}	R^2_{Bias}
OC	-.01	.02	.73	-.40	.02	.00	.88	.02	.00	.93	.16	.77
ocq.parcel1	-.01	.03	.87	-.35	.03	.00	.78	.03	.00	.73	.12	.60
ocq.parcel2	-.01	.03	.85	-.34	.03	.00	.80	.03	.00	.75	.11	.64
ocq.parcel3	-.01	.04	.86	-.26	.04	.00	.76	.04	.00	.65	.07	.58
ocq.parcel4	.00	.04	.95	-.39	.04	.00	.63	.04	.00	.56	.15	.40
ocq.parcel5	.00	.04	.92	-.37	.04	.00	.70	.04	.00	.63	.14	.49
SRK	-.02	.01	.08	.39	.01	.00	.96	.01	.00	.97	.15	.93
srk.parcel1	-.02	.02	.38	.40	.02	.00	.91	.02	.00	.88	.16	.82
srk.parcel2	-.02	.02	.37	.36	.02	.00	.93	.02	.00	.89	.13	.86
srk.parcel3	-.02	.02	.42	.35	.02	.00	.91	.02	.00	.86	.12	.84
srk.parcel4	-.02	.02	.42	.33	.02	.00	.92	.02	.00	.86	.11	.85
srk.parcel5	-.02	.02	.41	.42	.02	.00	.88	.02	.00	.85	.18	.78
gc	.01	.05	.86	.47	.05	.00	.20	.05	.00	.26	.23	.04
gc.parcel1	.01	.05	.88	.41	.05	.00	.14	.05	.01	.18	.16	.02
gc.parcel2	.01	.05	.90	.37	.05	.00	.16	.05	.00	.16	.13	.03
gc.parcel3	.01	.06	.90	.34	.06	.00	.12	.06	.04	.13	.12	.01
gc.parcel4	.01	.05	.89	.38	.05	.00	.18	.05	.00	.18	.15	.03
gc.parcel5	.01	.06	.90	.34	.06	.00	.04	.06	.43	.11	.11	.00
gf	.00	.06	.98	.27	.06	.00	.06	.06	.28	.07	.07	.00
DT	.00	.06	1.00	.24	.06	.00	.28	.06	.00	.14	.06	.08
Openness Creativity	.00	.06	1.00	.12	.06	.04	.26	.06	.00	.08	.01	.07
Honesty-Humility	.00	.06	1.00	.14	.06	.01	-.01	.06	.86	.02	.02	.00
Faking Ability	-.02	.06	.78	.26	.06	.00	.12	.07	.08	.07	.07	.01

Note. Each line represents one regression model. In each regression model, d' and c were used as predictors.

SM Figure 2

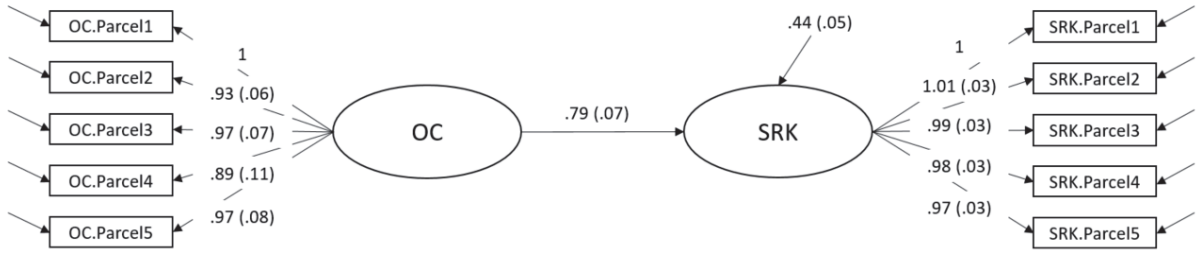
Confirmatory Factor Model of Faking Ability and Knowledge (Gc)



Note. All parameters are standardized – unstandardized parameter estimates are depicted on the right. $n = 298$; $\chi^2(34) = 32.40$, CFI = 1, RMSEA = .000, SRMR = .035.

SM Figure 3

Structural Model of Overclaiming (OC) and Self-Reported Knowledge (SRK)

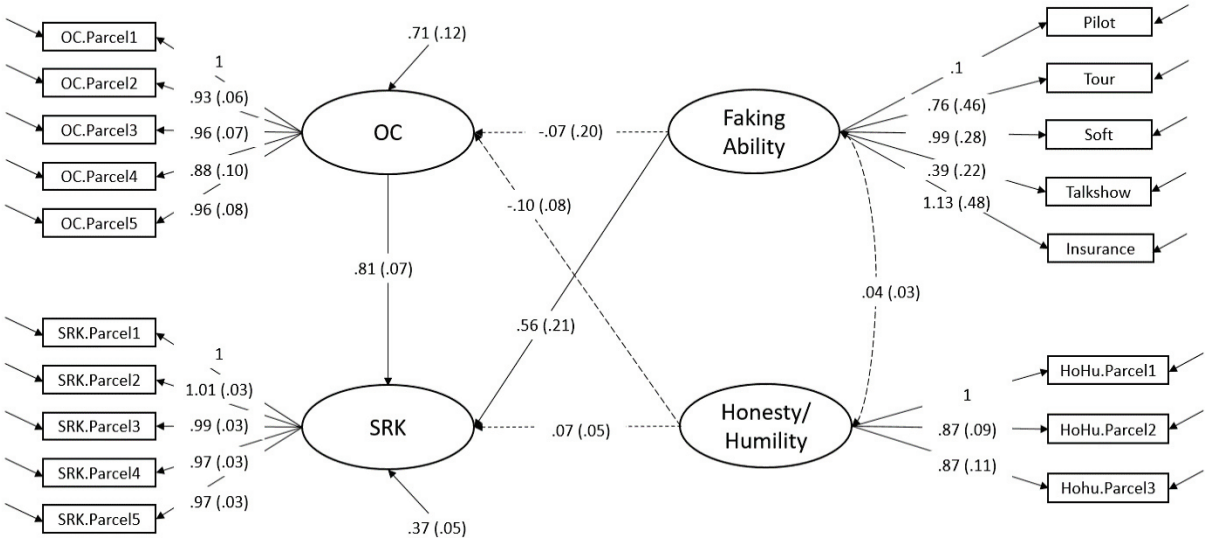


Note. All parameters are unstandardized. $n = 298$; $\chi^2(34) = 90.87$, CFI = .978, RMSEA = .075, SRMR = .038.

Although both factors share a substantial amount of variance, they both capture unique variance distinct to their indicators. We tested the model depicted in Figure 1 against a g-factor model with a chi-square difference test. The test suggested a significantly better fit of our here presented model compared to a g-factor model ($\Delta\chi^2(1, N = 298) = 176.23$, $p < 0.001$).

SM Figure 4

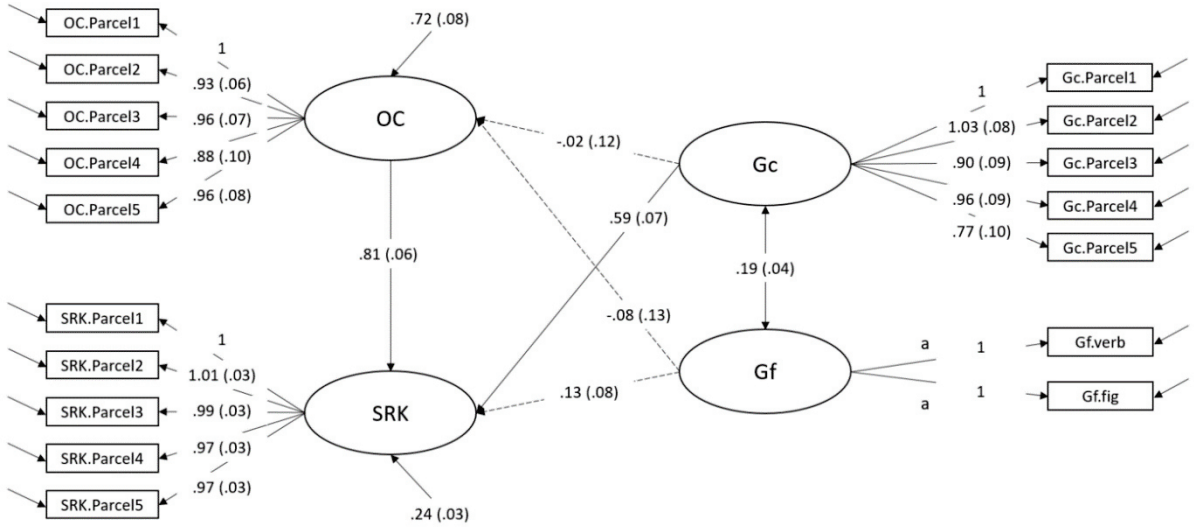
Structural Model of Overclaiming (OC), Self-Reported Knowledge (SRK), Honesty-Humility, and Faking Ability



Note. All parameters are unstandardized; non-significant coefficients are displayed with dashed lines ($p < .05$). n =298; $\chi^2(129) = 215.64$, CFI = .972, RMSEA= .047, SRMR = .049.

SM Figure 5

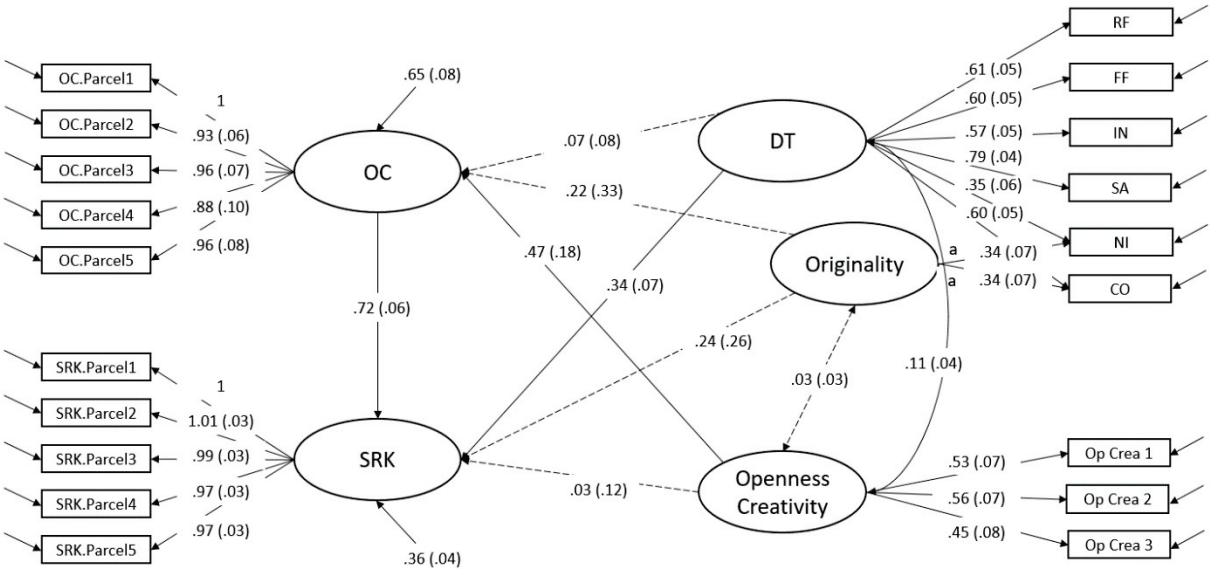
Structural Model of Overclaiming (OC), Self-Reported Knowledge (SRK), and Knowledge (Gc) and Reasoning Ability (Gf)



Note. All parameters are unstandardized; non-significant coefficients are displayed with dashed lines ($p < .05$). $n = 298$; $\chi^2(114) = 199.78$, CFI = .975, RMSEA = .050, SRMR = .040.

SM Figure 6

Structural Model of Overclaiming (OC), Self-Reported Knowledge (SRK), and Divergent Thinking, Originality, and Openness



Note. RF = Retrieval Fluency, FF = Figural Fluency, IN = Inventing Names, SA = Similar Attributes, NI = Nicknames, CO = Combining Objects. All parameters are standardized; non-significant coefficients are displayed with dashed lines ($p < .05$). $n = 298$; $\chi^2(142) = 225.38$, CFI = .975, RMSEA = .044, SRMR=.044.

The Nomological Net of Knowledge, Self-Reported Knowledge, and Overclaiming in Children

Supplementary Material

Method

Data Preparation and Parcel Building

Prior to the main statistical analyses, the data of each measurement instrument were prepared to derive meaningful indicators per scale. For the OCQ, we examined the mean familiarity ratings of all items and compared them across the two pseudo-parallel versions. There was no difference present, indicating that the reverse presentation order of the items had no effect on rating familiarity with the terms. We built parcels for the foils (indicating overclaiming) and the reals (indicating self-reported knowledge) based on the mean familiarity ratings of the single items after inspecting the corrected item-test correlations of the single items per subscale (see SM Table 1); the correlations with for all items were relatively homogeneous, indicating no abnormalities. We therefore decided to keep all items of the questionnaire for building parcels. The six parcels per scale were built randomly, but in a way that each parcel was stratified across all broad content domains (Cole et al., 2016), similar to other studies administering OCQs (Goecke et al., 2020; Hülür et al., 2011). Regarding the intelligence scales, we used the same approach as Schroeders and colleagues (2016) for identifying possibly problematic single items. For the Gf scales, we computed unidimensional measurement models based on the single dichotomous items across both parallel test versions, as the items of version B were isomorphs of version A. Based on these measurement models, we excluded two items of the verbal scale, due to considerably improved model fit without these two items. For the figural and numerical scale, no items were excluded. After establishing unidimensionality, a total mean score was built for each scale, resulting in altogether three indicators for Gf. For Gc, we first examined the test versions A and B

individually, as they shared no items. For both test versions, we considered the three subscales (i.e., broad knowledge domains) they were supposed to capture (humanities, social sciences, and natural sciences). For each subscale, we first examined the item difficulties and excluded items below the corresponding guessing probability (i.e., *guessing*; cutoff-value for considering an item answered with guessing probability was the upper boundary of the confidence interval around the statistical guessing probability; Schroeders et al., 2016). After that, we computed unidimensional measurement models per subscale and excluded items with non-significant factor loadings (*lambda*), or items which considerably deteriorated model fit (*misfit*). Subsequently to dropping items per subscale, we calculated total mean scores per subscale (i.e., knowledge domain) per test version. The scores were then z-standardized in order to account for diverging item contents of the test versions. Eventually, the standardized scores per test version were combined into one variable per subscale resulting in altogether three indicators for Gc. For the openness factor of personality, we dropped two items in order to ensure unidimensionality of the scale and then parceled according to item content (i.e., facets). For indicating epistemic curiosity and fluency, we used the raw items. Please see the following table (SM Table 2) for an overview of final unidimensional measurement models for each scale based on the single items.

Results

SM Table 1

Descriptive Statistics of All OCQ Items.

Real = 1, Foil = 2	Parcel	Item	Domain	*M*	*SD*	Skew	Kurtosis	$\Delta_{M\ \text{A vs. B}}$	r_{it}
1	1	Reagenzglas	Nat	1.8	1.33	1.51	.91	< .1	.47
1	2	Hormon	Nat	1.82	1.32	1.5	.88	< .1	.52
1	3	Kompass	Nat	3.97	1.31	-.92	-.42	< .1	.5
1	4	Organ	Nat	3.4	1.46	-.27	-1.32	< .1	.54
1	5	Kamera	Nat	4.35	1.15	-1.71	1.8	< .1	.4
1	6	Schwerkraft	Nat	3.55	1.48	-.47	-1.21	< .1	.57

2	1	Wasserstoffkabel	Nat	2.16	1.53	.96	-.68	> .1	.60
1	1	Magnet	Nat	4.3	1.33	-1.9	-.11	< .1	.46
1	2	Nordpol	Nat	4.17	1.18	-1.18	.29	< .1	.49
1	3	Regenwald	Nat	3.89	1.33	-.78	-.74	< .1	.54
2	2	Militis	Nat	1.55	1.21	2.12	3	< .1	.66
1	4	Staubgefäß	Nat	2.28	1.57	.81	-.94	< .1	.61
1	5	Quotient	Nat	1.59	1.23	1.97	2.42	< .1	.44
1	6	Molekül	Nat	2.12	1.44	1.1	-.42	< .1	.59
2	3	Reißfüßler	Nat	1.71	1.32	1.71	1.43	< .1	.69
1	1	Glasfaser	Nat	2.23	1.55	.88	-.83	< .1	.56
1	2	Allergie	Nat	3.83	1.37	-.77	-.73	< .1	.55
2	4	Sechsmeridian	Nat	1.49	1.14	2.28	3.84	< .1	.68
1	3	Atom	Nat	2.99	1.59	.9	-1.54	< .1	.57
1	4	Subtraktion	Nat	2.87	1.73	.13	-1.71	< .1	.49
1	5	Karies	Nat	3.5	1.71	-.4	-1.69	< .1	.48
1	6	Infrarot	Nat	2.5	1.51	1.1	-.41	< .1	.55
1	1	Labor	Nat	3.89	1.31	-.79	-.64	< .1	.56
1	2	Energie	Nat	3.99	1.28	-.91	-.47	< .1	.57
1	3	Ökosystem	Nat	2.34	1.54	.74	-1	< .1	.6
2	5	Bertlsche Probe	Nat	1.48	1.15	2.33	4	< .1	.68
1	4	Diagnose	Nat	2.64	1.53	.49	-1.24	< .1	.6
1	5	Formel	Nat	3.1	1.54	.6	-1.5	< .1	.58
1	6	Bronchitis	Nat	2.18	1.51	.91	-.73	< .1	.48
1	1	Atmospähre	Nat	2.73	1.63	.32	-1.51	< .1	.58
1	2	Satellit	Nat	3.31	1.54	-.24	-1.42	< .1	.58
1	3	Tornado	Nat	3.71	1.46	-.67	-.98	< .1	.55
2	6	Plomode	Nat	1.59	1.22	1.95	2.38	< .1	.67
2	1	Durchzahl	Nat	2.6	1.5	1.9	-.39	< .1	.63
1	4	Kristall	Nat	3.98	1.26	-.87	-.47	< .1	.54
1	5	Geometrie	Nat	3.12	1.58	-.2	-1.55	< .1	.6
1	6	Kondensator	Nat	1.74	1.33	1.63	1.16	< .1	.52
1	1	Lösungsmittel	Nat	3.14	1.55	-.1	-1.52	< .1	.59
1	2	Thermometer	Nat	3.72	1.5	-.71	-.99	< .1	.44
1	3	Steppe	Nat	2.36	1.55	.71	-1.6	< .1	.48
1	4	Statistik	Nat	2.16	1.46	.98	-.51	< .1	.55
1	5	Lawine	Nat	3.69	1.45	-.62	-1.1	< .1	.49
2	2	Sichelmotor	Nat	1.82	1.35	1.48	.73	< .1	.70
1	6	Therapie	Nat	3.4	1.6	.5	-1.57	< .1	.58
1	1	Logik	Hum	2.71	1.53	.4	-1.3	< .1	.61
1	2	Präsens	Hum	2.25	1.57	.86	-.89	< .1	.5
1	3	Melodie	Hum	3.87	1.39	-.86	-.62	< .1	.52
2	3	Versmeter	Hum	1.59	1.2	2.3	2.81	< .1	.69
1	4	Ethik	Hum	1.65	1.27	1.85	1.97	< .1	.53
2	4	Abbaute	Hum	1.66	1.26	1.85	1.99	< .1	.65
1	5	Oper	Hum	3.24	1.61	-.18	-1.53	> .1	.45

1	6	Fabel	Hum	2.39	1.61	.69	-1.16	< .1	.49
1	1	Tusche	Hum	2.33	1.57	.76	-1.1	< .1	.5
2	5	Gewissmus	Hum	1.78	1.32	1.56	1.2	< .1	.70
1	2	Atelier	Hum	1.99	1.46	1.17	-.19	< .1	.49
1	3	Budda	Hum	1.94	1.49	1.26	-.6	< .1	.48
1	4	Hip Hop	Hum	3.63	1.47	-.51	-1.22	< .1	.46
1	5	Bibliothek	Hum	4	1.42	-1.8	-.31	< .1	.43
1	6	Taufe	Hum	3.86	1.43	-.86	-.67	< .1	.45
1	1	Palette	Hum	2.57	1.64	.47	-1.41	< .1	.51
1	2	Epos	Hum	1.58	1.23	2.2	2.59	< .1	.53
2	6	Afrikanischer Walzer	Hum	2.11	1.48	1.3	-.46	< .1	.64
1	3	Roman	Hum	2.95	1.61	.14	-1.55	< .1	.56
1	4	Toleranz	Hum	1.88	1.4	1.42	.5	< .1	.59
1	5	Pinsel	Hum	4.41	1.11	-1.83	2.26	< .1	.36
1	6	Fantasie	Hum	4.19	1.32	-1.41	.58	< .1	.39
1	1	Papst	Hum	3.63	1.55	-.6	-1.2	< .1	.47
1	2	Fotografie	Hum	3.96	1.34	-.91	-.53	< .1	.53
1	3	Bewusstsein	Hum	3.29	1.56	-.21	-1.47	< .1	.6
1	4	Sonate	Hum	1.69	1.32	1.73	1.45	< .1	.51
1	5	Altar	Hum	3.72	1.46	-.67	-1.1	< .1	.48
1	6	Dirigent	Hum	3.14	1.67	-.11	-1.64	< .1	.5
2	1	Quergriffel	Hum	1.75	1.33	1.61	1.12	< .1	.65
1	1	Seele	Hum	3.51	1.51	-.44	-1.28	< .1	.56
1	2	Industrie	Soc	2.3	1.41	.84	-.58	< .1	.57
1	3	Börse	Soc	2.83	1.5	.29	-1.32	< .1	.56
1	4	Faustkeil	Soc	1.69	1.27	1.75	1.68	< .1	.49
1	5	Parlament	Soc	2.4	1.39	1.11	-.13	< .1	.56
1	6	Revolution	Soc	1.93	1.35	1.29	.31	< .1	.58
1	1	Bilanz	Soc	1.8	1.29	1.53	1	< .1	.52
1	2	Bußgeld	Soc	2.84	1.59	.22	-1.49	< .1	.51
1	3	Zinsen	Soc	2.87	1.5	.19	-1.37	< .1	.57
1	2	Amtsregler	Soc	1.92	1.38	1.28	.2	< .1	.59
1	4	Papyrus	Soc	2	1.48	1.16	-.25	< .1	.47
1	5	Demokratie	Soc	2.47	1.48	.66	-.97	< .1	.62
1	6	Kerker	Soc	3.65	1.51	-.6	-1.13	> .1	.47
1	1	Wettbewerb	Soc	4.24	1.19	-1.38	.76	> .1	.44
2	3	Friede von Leipzig	Soc	2.6	1.5	1.1	-.37	< .1	.62
1	2	Republik	Soc	3	1.6	.9	-1.55	< .1	.58
1	3	Kartell	Soc	1.65	1.25	1.86	2.8	< .1	.52
1	4	Gesetz	Soc	3.48	1.46	-.39	-1.22	< .1	.58
1	5	Präsident	Soc	3.68	1.44	-.6	-1.5	< .1	.55
1	6	Verfassung	Soc	2.54	1.48	.58	-1.6	< .1	.63
1	1	Kapital	Soc	2.56	1.59	.55	-1.28	< .1	.55
2	4	Dachgericht	Soc	1.85	1.38	1.44	.57	< .1	.68
1	2	Sparkasse	Soc	4.15	1.2	-1.14	.12	> .1	.47

1	3	Produktion	Soc	2.86	1.6	.19	-1.52	< .1	.63
2	5	Teilerkonto	Soc	1.95	1.42	1.29	.17	< .1	.65
1	4	Meineid	Soc	1.54	1.18	2.17	3.3	< .1	.47
1	5	Auktion	Soc	2.12	1.52	1	-.6	< .1	.54
1	6	Ministerium	Soc	2.37	1.5	.75	-.9	< .1	.62
1	1	Anwalt	Soc	3.3	1.51	-.18	-1.43	< .1	.59
1	2	Babylon	Soc	2.9	1.5	1.5	-.48	< .1	.51
2	6	Patalpakt	Soc	1.66	1.31	1.82	1.75	< .1	.68

Note. Nat = nature and technology, soc = politics and society, hum = art and culture. $\Delta_{M\ \text{A vs. B}}$ gives the unstandardized mean difference for an item between test version A and test version B. The reported item-test correlations (r_{it}) for the single items correspond to the respective subscales of the OCQ (OC, as indicated through the foils, and SRK, as indicated through the reals, respectively).

SM Table 2

Measurement Models for All Scales.

	Measurement Model	*n* (Indicators) before	*n* (Indicators) after	*n* (Persons)	χ^2	df	CFI	RMSEA	[90% CI]	SRMR	ω	exclusion due to
1	Gff	16	16	836	217.11	104	.951	.036	[.029; .043]	.074	.78	
2	Gfv	9	7	848	20.72	14	.959	.024	[.000; .044]	.049	.41	misfit (2 items)
3	Gfn	9	9	848	42.43	27	.973	.026	[.008; .040]	.054	.58	
4	Gc A nat	15	13	469	70.23	65	.987	.013	[.000; .031]	.064	.58	guessing (1 item), lambda (1 item)
5	Gc B nat	16	13	375	70.03	65	.962	.014	[.000; .035]	.075	.49	guessing (1 item), lambda (1 item), misfit (1 item)
6	Gc A hum	14	10	469	41.86	35	.958	.020	[.000; .041]	.061	.48	guessing (3 items), lambda (1 item)
7	Gc B hum	12	11	375	50.46	44	.975	.020	[.000; .041]	.068	.58	lambda (1 item)
8	Gc A soc	13	11	469	47.91	44	.982	.014	[.000; .035]	.060	.50	guessing (1 item), lambda (1 item)
9	Gc B soc	14	10	375	24.38	35	1	.000	[.000; .015]	.050	.55	guessing (2 items), lambda (2 items)
10	Honesty-Humility	10	8	578	83.53	20	.920	.074	[.061; .088]	.047	.81	misfit (2 items)
11	Openness	10	8	578	96.98	20	.911	.082	[.067; .097]	.048	.79	misfit (2 items)

Note. Gff = figural fluid intelligence; Gfv = verbal fluid intelligence; Gfn = numerical fluid intelligence; Gc = crystallized intelligence; nat = natural sciences; hum = humanities; soc = social sciences; A = version A of a test; B = version B of a test; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual; ω = McDonald's Omega (McDonald, 1999).

SM Table 3

Measurement Invariance Testing for a Correlated Factor Model with OC, SRK, and Gc.

Invariance Step	χ2	df	CFI	RMSEA	SRMR	Δ(χ^2)	Δ(*df*)	Δ(CFI)
Configural	696.27	174	.960	.083	.043			
Metric	701.30	186	.960	.080	.044	4.90	12	0
Scalar	743.32	198	.958	.079	.046	42.25	12	.002
Strict	775.89	213	.957	.078	.046	19.51	15	.001

Note. *n*(boys) = 447, *n*(girls) = 423

Signal Detection Analysis

Following previous scoring procedures of OCQs (e.g., Paulhus & Petrusic, 2007), we dichotomized participants' responses not only at the lowest cut-off of the response-scale (i.e., 1 = unfamiliar; > 1 = familiar"), but across all cut-offs in order to stabilize the statistical indices (i.e., 1 & 2 = unfamiliar; > 2 = familiar" etc.). Four OC accuracy scores and four OC bias scores were computed and averaged to obtain signal detection parameter estimates (please see SM Figure 1 for a full correlation plot of the indices). For each cut-off, we calculated the *Hit-Ratio* (HR) and the *False-Alarm-Ratio* (FAR) as follows after dichotomizing:

$$HR = \frac{Hits + 0.5}{Signal\ Trials + 1} \qquad (1)$$

$$FAR = \frac{False\ Alarms + 0.5}{Noise\ Trials + 1} \qquad (2)$$

The formulas of the ratios were adjusted to ensure that indices are corrected for extreme hit-rates and extreme false-alarm-rates (i.e., 0 or 1; see Hautus, 1995). The following two indices were calculated within the framework of signal detection analysis. The *accuracy index d'* (d prime) and the response bias *index c* (criterion location, see also Macmillan, 1993; Stanislaw & Todorov, 1999):

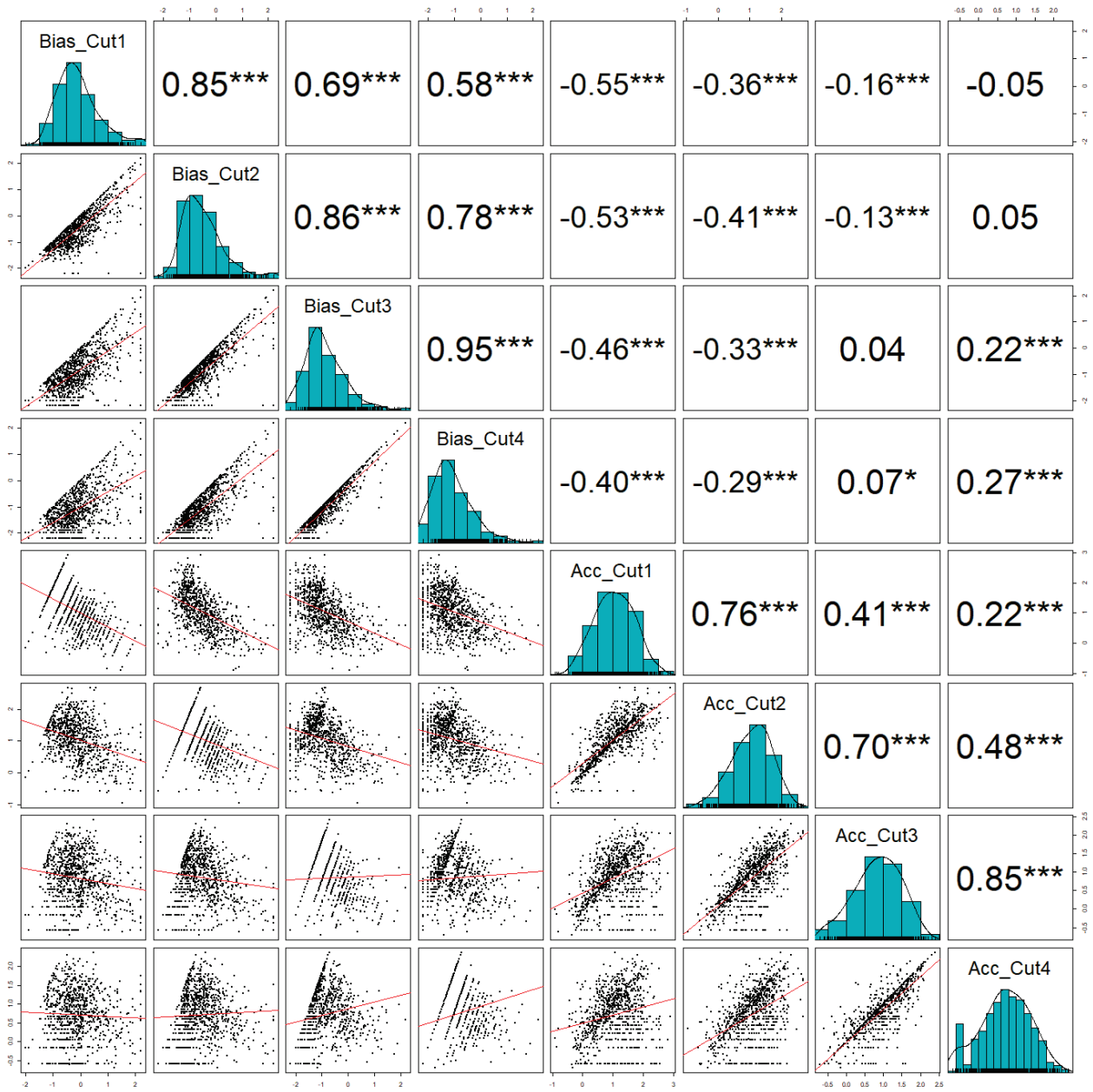
$$d' = \Phi^{-1}(HR) - \Phi^{-1}(FAR) \qquad (3)$$

$$c = \frac{\Phi^{-1}(HR) + \Phi^{-1}(FAR)}{2} \qquad (4)$$

Please note that the correlation between OC accuracy and OC bias was -.25 across both test versions A and B in our sample (r = -.22 and -.29 for the single test versions, respectively). Therefore, the magnitude of this particular correlation fits into the hitherto heterogeneous picture of associations between OC accuracy and OC bias found across the literature (*r* = -.01, *r* = .15, Dunlop et al., 2017, 2019; *r* = -.01, Goecke et al., 2020; *r* = -.21, Paulhus et al., 2003; *r* = .26 and r = -.74, Ziegler et al., 2013). The results for the signal detection approach can be found in SM Table 4.

SM Figure 1

Correlation Plot of OC Accuracy and OC Bias Scores for All Four Cut-Offs Across Both Test Versions



Note. Distributions are depicted on the diagonal. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

SM Table 4

Regression Models With d' (OC Accuracy) and c (OC bias) As Predictors.

Criterion	Intercept			OC Accuracy (*d'*)			OC Bias (*c*)					
	B	SE_B	p_B	β_{Acc}	SE_{Acc}	p_{Acc}	β_{Bias}	SE_{Bias}	p_{Bias}	R^2	R^2_{Acc}	R^2_{Bias}
OC	.00	.01	1.00	-.34	.01	.00	.84	.01	.00	.96	.11	.71
ocq.parcel1	.00	.02	.89	-.28	.02	.00	.72	.02	.00	.69	.08	.52
ocq.parcel2	.00	.02	1.00	-.27	.02	.00	.76	.02	.00	.74	.07	.57
ocq.parcel3	.00	.02	1.00	-.29	.02	.00	.75	.02	.00	.75	.08	.56
ocq.parcel4	.00	.02	1.00	-.32	.02	.00	.73	.02	.00	.75	.11	.53
ocq.parcel5	.00	.02	.96	-.30	.02	.00	.75	.02	.00	.76	.09	.56
ocq.parcel6	.00	.02	.95	-.31	.02	.00	.77	.02	.00	.81	.10	.59
SRK	.00	.01	1.00	.32	.01	.00	1.00	.01	.00	.93	.10	.99
srk.parcel1	.00	.01	1.00	.29	.01	.00	.94	.01	.00	.84	.09	.89
srk.parcel2	.00	.01	1.00	.34	.02	.00	.93	.02	.00	.82	.11	.87
srk.parcel3	.00	.01	1.00	.34	.02	.00	.93	.02	.00	.81	.12	.86
srk.parcel4	.00	.01	1.00	.12	.01	.00	.94	.01	.00	.84	.01	.89
srk.parcel5	.00	.01	1.00	.44	.01	.00	.91	.01	.00	.83	.20	.83
srk.parcel6	.00	.02	1.00	.27	.02	.00	.93	.02	.00	.81	.07	.86
Gff	.00	.03	.98	.21	.04	.00	-.11	.04	.00	.07	.04	.01
Gff.parcel1	.00	.03	.99	.15	.04	.00	-.03	.04	.34	.03	.02	.00
Gff.parcel2	.00	.03	.99	.16	.04	.00	-.13	.04	.00	.05	.02	.02
Gff.parcel3	.00	.03	.98	.16	.04	.00	-.06	.04	.12	.03	.03	.00
Gff.parcel4	.00	.03	.99	.16	.04	.00	-.11	.04	.00	.05	.03	.01
Gfn	.00	.03	1.00	.26	.03	.00	.01	.03	.83	.07	.07	.00
Gfn.parcel1	.00	.03	1.00	.18	.04	.00	.00	.04	.94	.03	.03	.00
Gfn.parcel2	.00	.03	1.00	.19	.04	.00	.02	.04	.49	.03	.04	.00
Gfn.parcel3	.00	.03	1.00	.21	.04	.00	-.01	.04	.72	.05	.05	.00
Gfv	.00	.03	1.00	.26	.03	.00	.02	.03	.50	.06	.07	.00
Gfv.parcel1	.00	.03	1.00	.13	.04	.00	-.03	.04	.35	.02	.02	.00
Gfv.parcel2	.00	.03	1.00	.18	.04	.00	.02	.04	.63	.03	.03	.00
Gfv.parcel3	.00	.03	.99	.22	.04	.00	.06	.04	.09	.04	.05	.00
Gc_soc.parcel	.00	.03	.98	.32	.03	.00	.07	.03	.04	.09	.10	.01
Gc_sci.parcel	.00	.03	.98	.33	.03	.00	.08	.03	.02	.10	.11	.01
Gc_hum.parcel	.00	.03	.98	.34	.03	.00	.08	.03	.02	.11	.12	.01
Fluency	.00	.07	1.00	.05	.07	.49	.13	.07	.07	.02	.00	.02
EC	.00	.07	1.00	.13	.07	.08	.23	.07	.00	.06	.02	.05
EC Interest	.00	.07	1.00	.09	.07	.23	.14	.07	.05	.03	.01	.02
EC Deprivation	.00	.07	1.00	.15	.07	.03	.28	.07	.00	.09	.02	.08
Openness	.00	.04	.98	.12	.04	.01	.10	.05	.03	.02	.02	.01
Honesty-Humility	.00	.04	1.00	-.05	.04	.26	-.10	.04	.02	.01	.00	.01

Note. Each line represents one regression model. In each regression model, d' (OC Accuracy) and c (OC Bias) were used as predictors.

References

Cole, D. A., Perkins, C. E., & Zelkowitz, R. L. (2016). Impact of homogeneous and heterogeneous parceling strategies when latent variables represent multidimensional constructs. *Psychological Methods*, *21*, 164–174. https://doi.org/10.1037/met0000047

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

Dunlop, P. D., Bourdage, J. S., de Vries, R. E., McNeill, I. M., Jorritsma, K., Orchard, M., Austen, T., Baines, T., & Choe, W.-K. (2019). Liar! Liar! (when stakes are higher): Understanding how the overclaiming technique can be used to measure faking in personnel selection. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000463

Goecke, B., Weiss, S., Steger, D., Schroeders, U., & Wilhelm, O. (2020). Testing competing claims about overclaiming. *Intelligence*, *81*, 101470. https://doi.org/10.1016/j.intell.2020.101470

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values ofd′. *Behavior Research Methods, Instruments, & Computers*, *27*, 46–51. https://doi.org/10.3758/BF03203619

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

Macmillan, N. A. (1993). Signal detection theory as data analysis method and psychological decision model. In G. Keren & C. Lewis (Hrsg.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (S. 21–57). Erlbaum Associates, Inc.

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

Paulhus, D. L., & Petrusic, W. M. (2007). Measuring individual differences with signal detection analysis: A guide to indices based on knowledge ratings. *Unpublished Manuscript*.

Schroeders, U., Schipolowski, S., Zettler, I., Golle, J., & Wilhelm, O. (2016). Do the smart get smarter? Development of fluid and crystallized intelligence in 3rd grade. *Intelligence*, *59*, 84–95. https://doi.org/10.1016/j.intell.2016.08.003

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*, 137–149. https://doi.org/10.3758/BF03207704

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The Vocabulary and Overclaiming Test (VOC-T). *Journal of Individual Differences*, *34*, 32–40. https://doi.org/10.1027/1614-0001/a000093

Stop Worrying About Multiple-Choice: Fact Knowledge Does Not Change With Response Format

Supplementary Material

Method

Pilot Study

Please note that some of the items were used in a pilot study, where we sought to obtain some first difficulty estimates for an open-ended and a cued response format for the single items. In this pilot study, we tested $N = 161$ participants ($n = 33$ students and $n = 128$ *Prolific* members; $M_{age} = 32.78$). In the pilot study, we administered 60 items of the item pool from Steger and colleagues (2019) covering the same broad knowledge domains (natural sciences, social sciences, life sciences, and humanities). However, not all participants worked on all items, or on all response formats, respectively. Prior to responding to an item, participants were presented only with the knowledge question and were asked to indicate whether they knew the answer, whether they did not know the answer, or whether they would know the answer if they were presented with a cue. Participants stating that they would know the answer to a question were referred to the item in the open-ended response format, whereas participants stating they might know the answer if presented with a cue were referred to the cued open-ended item format. Participants stating not to know an answer were referred to an MC response format and instructed to guess. The number of participants working on a specific item in the cued open-ended item format ranged from $n = 9$-36 and in the open-ended format from $n = 3$-126. After the pilot study, we replaced some of the items, which were either too easy (i.e., all participants solved it correctly) or too difficult (i.e., no participants solved it correctly). Additionally, we decided to add more items to our scale for the main study to obtain a broader assessment of the construct.

Results

SM Table 1

Descriptive Statistics of All Gc Items.

	MC					Cued				Open			
	Steger et al. (2019)	Study 1 (*N*=46-50)		Study 2 (*N* = 300)		Study 1 (*N*=46-51)		Study 2 (*N* = 300)		Study 1 (*N*=45-50)		Study 2 (*N* = 300)	
Item	*M*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
bio013	0.72	0.48	0.50			0.13	0.34			0.04	0.21	0.05	0.23
bio036	0.33	0.26	0.44	0.39	0.49	0.16	0.37			0.22	0.42		
bio067	0.47	0.33	0.47			0.87	0.34	0.92	0.27	0.56	0.50		
bio068	0.92	0.76	0.43			0.72	0.46			0.59	0.50	0.67	0.47
bio083	0.56	0.52	0.50			0.33	0.47			0.20	0.40	0.16	0.37
bio084	0.87	0.78	0.42	0.75	0.43	0.64	0.48			0.43	0.50		
bio088	0.78	0.65	0.48			0.20	0.40	0.22	0.42	0.12	0.33		
che014	0.36	0.44	0.50			0.09	0.28			0.02	0.15	0.02	0.15
che023	0.48	0.35	0.48	0.50	0.50	0.22	0.42			0.11	0.31		
che031	0.99	0.85	0.36			0.61	0.49	0.60	0.49	0.56	0.50		
che035	0.95	0.80	0.40	0.88	0.33	0.26	0.44			0.24	0.43		
che050	0.41	0.54	0.50	0.57	0.50	0.34	0.48			0.26	0.44		
che072	0.45	0.36	0.48			0.63	0.49			0.22	0.42	0.21	0.41
che101	0.49	0.46	0.50			0.24	0.43	0.26	0.44	0.24	0.43		
nut053	0.3	0.33	0.47			0.17	0.38	0.11	0.32	0.12	0.33		
nut074	0.99	1.00	0.00	0.99	0.10	0.92	0.27			0.67	0.47		
nut081	0.45	0.37	0.49	0.39	0.49	0.16	0.37			0.07	0.25		
nut089	0.48	0.46	0.50			0.30	0.47			0.26	0.44	0.34	0.47
hea008	0.57	0.50	0.51			0.35	0.48			0.39	0.49	0.31	0.46
hea056	0.99	0.96	0.20			0.54	0.50			0.65	0.48	0.69	0.46
hea057	0.6	0.41	0.50			0.33	0.47	0.35	0.48	0.20	0.40		
hea063	0.79	0.61	0.49			0.48	0.51	0.54	0.50	0.42	0.50		
hea064	0.97	0.96	0.21	0.95	0.21	0.86	0.35			0.87	0.34		
hea091	0.94	0.94	0.24			0.78	0.42			0.70	0.47	0.79	0.41
law012	0.84	0.35	0.48			0.22	0.42	0.08	0.27	0.20	0.40		
law015	0.56	0.26	0.44			0.39	0.49	0.34	0.47	0.12	0.33		
law029	0.92	0.89	0.31	0.91	0.29	0.68	0.47			0.50	0.51		
law068	0.84	0.74	0.44			0.43	0.50			0.37	0.49	0.32	0.47
law081	0.36	0.26	0.44	0.38	0.49	0.02	0.14			0.13	0.34		
law105	0.35	0.50	0.51			0.61	0.49	0.71	0.46	0.18	0.39		
art007	0.43	0.26	0.44			0.15	0.36	0.20	0.40	0.16	0.37		
art008	0.35	0.13	0.34	0.31	0.46	0.22	0.42			0.22	0.42		
art017	0.53	0.57	0.50	0.54	0.50	0.12	0.33			0.17	0.38		
art018	0.57	0.39	0.49			0.09	0.28	0.17	0.37	0.18	0.39		
art041	0.92	0.65	0.48			0.36	0.48	0.57	0.50	0.22	0.42		
art047	0.87	0.91	0.28			0.63	0.49			0.56	0.50	0.25	0.43
art028	0.88	0.52	0.50	0.65	0.48	0.24	0.43			0.20	0.40		
lit024	0.59	0.46	0.50	0.49	0.50	0.28	0.45			0.22	0.42		

lit040	0.52	0.44	0.50			0.41	0.50			0.49	0.51	0.48	0.50
lit042	0.88	0.57	0.50			0.33	0.47	0.44	0.50	0.20	0.40		
lit046	0.93	0.70	0.47	0.83	0.37	0.30	0.46			0.46	0.50		
lit051	0.95	0.62	0.49			0.67	0.47			0.17	0.38	0.22	0.42
lit124	0.36	0.34	0.48			0.46	0.50			0.07	0.25	0.15	0.36
med002	0.3	0.32	0.47			0.17	0.38			0.13	0.34	0.22	0.42
med009	0.54	0.43	0.50	0.45	0.50	0.42	0.50			0.41	0.50		
med047	0.85	0.72	0.46	0.77	0.42	0.28	0.45			0.48	0.51		
med060	0.44	0.20	0.40			0.15	0.36	0.23	0.42	0.14	0.35		
med062	0.9	0.87	0.34			0.61	0.49	0.73	0.44	0.46	0.50		
med078	0.63	0.57	0.50			0.09	0.28	0.02	0.15	0.02	0.14		
med093	0.57	0.39	0.49	0.49	0.50	0.74	0.44			0.35	0.48		
med096	0.65	0.46	0.50			0.57	0.50			0.39	0.49	0.33	0.47
mus015	0.45	0.22	0.42	0.30	0.46	0.04	0.20			0.00	0.00		
mus021	0.39	0.28	0.45			0.52	0.51			0.28	0.46	0.18	0.38
mus040	0.45	0.43	0.50			0.72	0.46	0.68	0.47	0.32	0.47		
mus043	0.62	0.38	0.49			0.72	0.46			0.43	0.50	0.44	0.50
mus046	0.84	0.83	0.38			0.11	0.31	0.27	0.44	0.12	0.33		
phy057	0.44	0.41	0.50			0.28	0.46	0.38	0.49	0.30	0.46		
phy075	0.81	0.56	0.50			0.24	0.43			0.35	0.48	0.26	0.44
phy090	0.56	0.43	0.50	0.46	0.50	0.08	0.27			0.22	0.42		
phy109	0.53	0.78	0.42			0.89	0.31	0.90	0.30	0.42	0.50		
pol029	0.46	0.48	0.50			0.33	0.47			0.28	0.46	0.26	0.44
pol034	0.73	0.63	0.49			0.59	0.50	0.64	0.48	0.35	0.48		
pol048	0.79	0.72	0.45			0.63	0.49			0.54	0.50	0.57	0.50
pol058	0.65	0.41	0.50	0.52	0.50	0.30	0.46			0.20	0.40		
pol070	0.32	0.28	0.46	0.24	0.43	0.10	0.31			0.07	0.25		
pol104	0.49	0.38	0.49			0.13	0.34			0.09	0.28	0.08	0.28
eco029	0.53	0.30	0.46			0.00	0.00			0.02	0.15	0.04	0.20
eco038	0.72	0.63	0.49			0.02	0.15	0.02	0.14	0.00	0.00		
eco050	0.89	0.83	0.38	0.87	0.34	0.46	0.50			0.50	0.51		
eco062	0.84	0.68	0.47			0.50	0.51			0.39	0.49	0.49	0.50
eco086	0.88	0.74	0.44			0.70	0.47	0.80	0.40	0.66	0.48		
eco090	0.48	0.52	0.51	0.54	0.50	0.10	0.30			0.30	0.47		

Note. bio = biology, che = chemistry, nut = nutrition, hea = health, law = law, art = art, lit = literature, med = medicine, mus = music, phy = physics, pol = politics, eco = economics. Biology, chemistry, and physics belong to the broad knowledge domain of natural sciences. Economics and politics belong to the broad knowledge domain of social sciences. Literature, art, and music belong to the broad knowledge domain of humanities. Nutrition, medicine, and health belong to the broad knowledge domain of life sciences.

SM Table 2

Odds Ratios for All Items (Study 1).

Item	MC vs. Cued			Cued vs. Open			MC vs. Open		
	OR	95%-CI		OR	95%-CI		OR	95%-CI	
		lower	upper		lower	upper		lower	upper
bio013	0.17	0.06	0.45	0.32	0.04	1.54	0.05	0.01	0.2
bio036	0.55	0.19	1.49	1.45	0.51	4.24	0.79	0.29	2.09
bio067*	13.08	4.78	41.4	0.2	0.06	0.53	2.59	1.14	6.11
bio068	0.8	0.32	2.03	0.56	0.23	1.35	0.45	0.18	1.09
bio083	0.45	0.19	1.03	0.51	0.19	1.32	0.23	0.09	0.56
bio084	0.5	0.19	1.23	0.44	0.19	0.99	0.22	0.08	0.54
bio088	0.13	0.05	0.34	0.57	0.17	1.75	0.08	0.02	0.21
che014	0.13	0.03	0.38	0.26	0.01	1.97	0.03	0	0.17
che023	0.53	0.21	1.32	0.44	0.13	1.36	0.24	0.07	0.69
che031	0.29	0.1	0.76	0.82	0.36	1.86	0.23	0.08	0.61
che035	0.09	0.03	0.23	0.9	0.35	2.29	0.08	0.03	0.21
che050	0.44	0.19	1	0.69	0.28	1.67	0.3	0.12	0.72
che072	2.98	1.31	7.03	0.17	0.06	0.41	0.5	0.19	1.23
che101	0.38	0.15	0.92	1	0.39	2.62	0.38	0.15	0.91
ern053	0.44	0.16	1.17	0.78	0.25	2.41	0.34	0.12	0.93
ern074	NA	NA	NA	0.19	0.05	0.58	NA	NA	NA
ern081	0.33	0.12	0.86	0.38	0.07	1.45	0.13	0.03	0.42
ern089	0.52	0.22	1.2	0.81	0.32	2.03	0.42	0.17	0.99
gsu008	0.54	0.23	1.22	1.2	0.51	2.85	0.65	0.28	1.46
gsu056	0.05	0.01	0.21	1.56	0.67	3.69	0.08	0.01	0.33
gsu057	0.69	0.29	1.63	0.52	0.2	1.32	0.36	0.14	0.89
gsu063	0.59	0.26	1.36	0.79	0.35	1.79	0.47	0.2	1.06
gsu064	0.3	0.04	1.35	1.08	0.32	3.71	0.32	0.04	1.54
gsu091	0.24	0.05	0.87	0.64	0.24	1.65	0.15	0.03	0.53
jur012	0.53	0.2	1.33	0.9	0.33	2.47	0.47	0.18	1.19
jur015	1.8	0.74	4.49	0.26	0.09	0.68	0.47	0.16	1.31
jur029	0.27	0.08	0.77	0.48	0.2	1.09	0.13	0.04	0.36

jur068	0.28	0.11	0.64	0.77	0.33	1.77	0.21	0.09	0.5
jur081*	0.07	0	0.37	6.5	1.01	173.35	0.43	0.14	1.26
jur105	1.55	0.67	3.59	0.15	0.05	0.36	0.23	0.08	0.56
kun007	0.52	0.17	1.45	1.22	0.41	3.77	0.63	0.23	1.68
kun008	1.85	0.63	5.95	0.99	0.36	2.64	1.82	0.6	5.96
kun017	0.11	0.04	0.29	1.53	0.48	5.14	0.17	0.06	0.42
kun018	0.16	0.04	0.47	2.55	0.77	10.25	0.4	0.15	0.97
kun028	0.31	0.13	0.7	0.5	0.19	1.23	0.15	0.06	0.38
kun041	0.17	0.04	0.52	0.75	0.33	1.71	0.13	0.03	0.38
kun047	0.3	0.12	0.7	0.78	0.28	2.13	0.23	0.09	0.56
lit024	0.47	0.2	1.09	0.72	0.27	1.84	0.34	0.13	0.83
lit040	0.9	0.39	2.03	1.41	0.62	3.27	1.27	0.56	2.87
lit042	0.38	0.16	0.88	0.52	0.2	1.32	0.2	0.08	0.48
lit046	0.19	0.08	0.45	1.94	0.84	4.59	0.37	0.15	0.87
lit051	1.26	0.54	2.97	0.11	0.04	0.27	0.13	0.05	0.34
lit124	1.62	0.71	3.75	0.12	0.03	0.36	0.19	0.05	0.59
med002	0.45	0.16	1.18	0.72	0.21	2.3	0.33	0.11	0.9
med009	0.94	0.42	2.14	0.97	0.43	2.21	0.92	0.4	2.11
med047	0.16	0.06	0.38	2.33	1	5.56	0.37	0.15	0.86
med060	0.74	0.24	2.23	1.06	0.34	3.35	0.79	0.26	2.3
med062	0.24	0.08	0.66	0.55	0.24	1.24	0.13	0.04	0.35
med078	0.08	0.02	0.23	0.24	0.01	1.81	0.02	0	0.1
med093	4.33	1.85	10.65	0.19	0.08	0.45	0.83	0.35	1.96
med096	1.52	0.68	3.45	0.5	0.21	1.15	0.76	0.33	1.71
mus015	NA	NA	NA	NA	NA	NA	NA	NA	NA
mus021	2.76	1.19	6.62	0.37	0.15	0.86	1.01	0.41	2.5
mus040	3.24	1.37	7.96	0.19	0.08	0.45	0.62	0.26	1.42
mus043	4.05	1.74	9.9	0.31	0.13	0.73	1.25	0.55	2.86
mus046	0.03	0.01	0.09	1.11	0.3	4.26	0.03	0.01	0.09
phy057	0.56	0.23	1.35	1.09	0.45	2.67	0.61	0.26	1.43
phy075	0.25	0.1	0.6	1.68	0.68	4.3	0.42	0.18	0.96
phy090	0.12	0.03	0.36	3.09	0.93	12.46	0.37	0.14	0.9

phy109	2.23	0.71	7.93	0.09	0.03	0.26	0.21	0.08	0.5
pol029	0.53	0.23	1.21	0.82	0.33	2.01	0.43	0.18	1
pol034	0.84	0.36	1.95	0.4	0.17	0.91	0.34	0.14	0.76
pol048	0.67	0.28	1.59	0.7	0.3	1.62	0.47	0.2	1.09
pol058	0.61	0.26	1.43	0.57	0.21	1.47	0.35	0.13	0.89
pol070	0.35	0.11	1.01	0.53	0.1	2.2	0.19	0.04	0.65
pol104	0.25	0.08	0.68	0.65	0.15	2.5	0.16	0.04	0.49
wir029	NA	NA	NA	NA	NA	NA	NA	NA	NA
wir038	0.02	0	0.08	0.92	0.02	36.57	0.01	0	0.07
wir050	0.18	0.07	0.46	1.17	0.52	2.64	0.22	0.08	0.55
wir062	0.48	0.2	1.09	0.65	0.28	1.48	0.31	0.13	0.71
wir086	0.81	0.32	2.03	0.85	0.35	2.02	0.69	0.28	1.67
wir090*	0.11	0.03	0.3	3.82	1.3	13.11	0.41	0.17	0.95

Note. Items marked with a * were excluded from analyses, because scoring the responses to these items were ambiguous. If values are NA, then the computation of odds ratios was not possible due to extreme responses in one of the two corresponding item formats (i.e., either all participants solved an item, or no participant solved an item). bio = biology, che = chemistry, nut = nutrition, hea = health, law = law, art = art, lit = literature, med = medicine, mus = music, phy = physics, pol = politics, eco = economics. Biology, chemistry, and physics belong to the broad knowledge domain of natural sciences. Economics and politics belong to the broad knowledge domain of social sciences. Literature, art, and music belong to the broad knowledge domain of humanities. Nutrition, medicine, and health belong to the broad knowledge domain of life sciences.

SM Table 3

Mean differences between test scores per response format of the single itemsets in Study 1.

Itemset	Response Formats X	Response Formats Y	*Mean (SD)* X	*Mean (SD)* Y	*t*	*df*	*p*	*d*	*d* 95%-CI
	MC	Cued	.54 (.16)	.39 (.18)	4.2	88.7	< .001	.87	[.43; 1.30]
A	MC	Open	.54 (.16)	.30 (.20)	6.3	84.95	< .001	1.3	[.85; 1.77]
	Cued	Open	.39 (.18)	.30 (.20)	2.3	88.6	< .05	.48	[.06; .90]
	MC	Cued	.53 (.16)	.34 (.19)	5.4	96.35	< .001	1.09	[.66; 1.51]
B	MC	Open	.53 (.16)	.28 (.17)	7.4	97.68	< .001	1.48	[1.03; 1.93]
	Cued	Open	.34 (.19)	.28 (.17)	1.6	97.46	.10	.33	[-.07; .72]
	MC	Cued	.54 (.16)	.41 (.19)	3.5	88.13	< .001	.72	[.29; 1.14]
C	MC	Open	.54 (.16)	.30 (.19)	6.3	87.76	< .001	1.32	[.86; 1.78]
	Cued	Open	.41 (.19)	.30 (.19)	2.7	89.98	< .01	.57	[.14; .99]

Note. d = Cohen's d. All provided descriptive statistics in this table refer to Figure 1 of the manuscript.

SM Table 4

Means, standard deviations, skew, kurtosis, and bivariate correlations for all indicators used in the measurement models of Study 2.

	Variable	*M*	*SD*	Skew	Kurtosis	1	2	3	4	5	6	7	8	9	10	11	12
1	gc.mc_nat	.59	.24	-.17	-.58	1											
2	gc.mc_soc	.71	.24	-.51	-.31	.27***	1										
3	gc.mc_hum	.52	.24	.23	-.66	.35***	.36***	1									
4	gc.mc_life	.53	.26	.00	-.65	.28***	.28***	.31***	1								
5	gc.open_nat	.23	.18	.72	.17	.30***	.32***	.29***	.37***	1							
6	gc.open_soc	.34	.26	.41	-.79	.30***	.46***	.34***	.31***	.37***	1						
7	gc.open_hum	.31	.25	.48	-.66	.31***	.38***	.53***	.35***	.37***	.45***	1					
8	gc.open_life	.40	.28	.27	-.69	.31***	.33***	.41***	.43***	.40***	.40***	.42***	1				
9	gc.cued_nat	.67	.21	-.06	-.70	.29***	.27***	.26***	.23***	.31***	.26***	.35***	.37***	1			
10	gc.cued_soc	.43	.21	-.12	-.70	.31***	.42***	.32***	.30***	.31***	.56***	.40***	.39***	.29***	1		
11	gc.cued_hum	.39	.25	.39	-.38	.30***	.31***	.56***	.38***	.33***	.39***	.53***	.49***	.31***	.47***	1	
12	gc.cued_life	.33	.23	.29	-.84	.34***	.27***	.39***	.46***	.48***	.41***	.42***	.56***	.37***	.43***	.49***	1
13	Defocusing	.07	.12	2.31	5.03	.20***	.14*	.17**	.13*	.27***	.27***	.27***	.13*	.20***	.21***	.17**	.23***

Note. *** $p < .001$; ** $p < .01$; * $p < .05$

SM Table 5

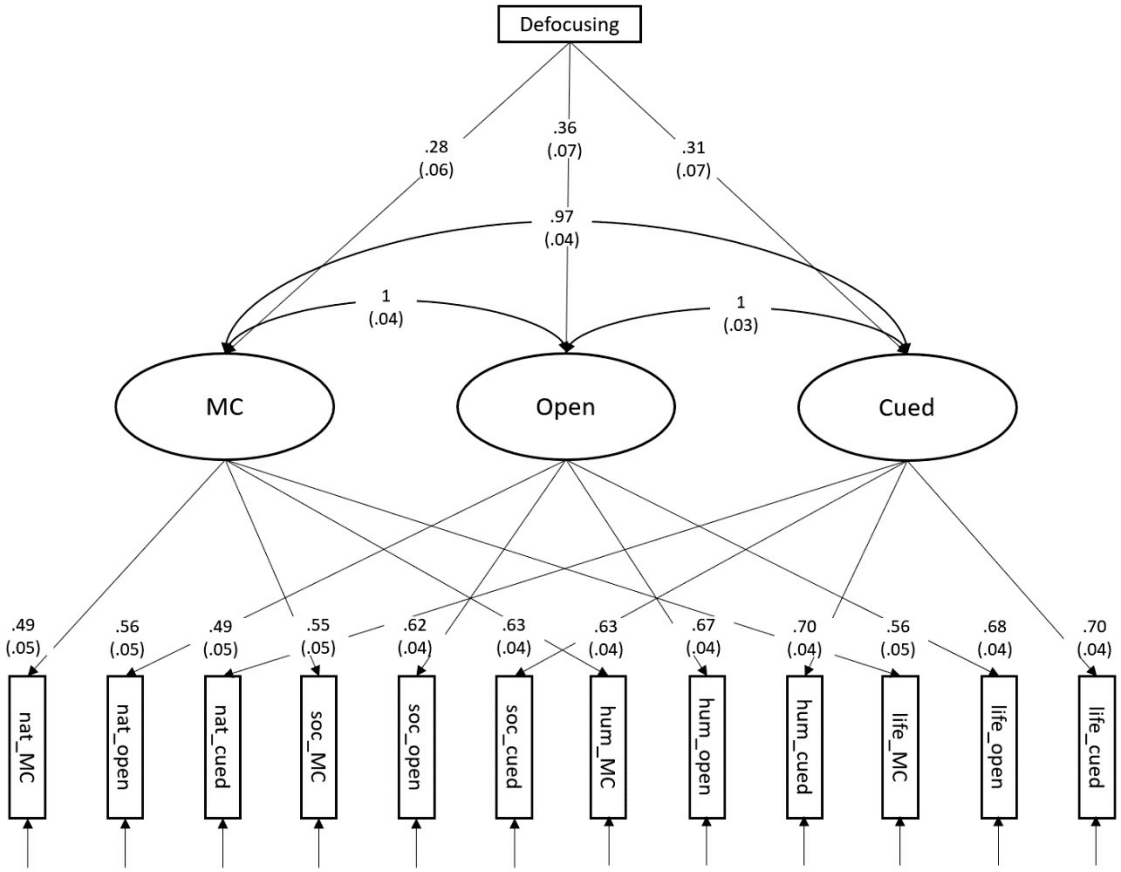
Measurement Models of the Response Formats.

	Measurement Model	*n* (Indicators)	*n* (Persons)	χ^2	df	CFI	RMSEA	[90% CI]	SRMR	ω	exclusion due to
1a	MC Items	23	308	382.94	230	.659	.047	[.038; .055]	.121	.65	guessing (1 item)
1b	MC Items	19	308	182.6	152	.920	.026	[.004; .038]	.093	.67	misfit (2 items), low r_{it} (2 items)
2a	Open Items	24	308	336.97	252	.907	.033	[.023; .042]	.118	.78	
2b	Open Items	22	308	261.97	209	.940	.029	[.016; .039]	.109	.79	misfit (1 item), low r_{it} (1 item)
3a	Cued Items	24	308	352.03	252	.915	.036	[.027; .045]	.132	.81	
3b	Cued Items	23	308	305.45	230	.932	.033	[.022; .042]	.129	.80	misfit (1 item)

Note. ω = McDonald's Omega (McDonald, 1999). Please note that one item for the MC scale had to be excluded from the scale initially, because the item was below guessing probability and had an $r_{it} < .05$. The unidimensional measurement model did not converge with the item included. We excluded MC items below the corresponding guessing probability for the MC format (i.e., *guessing*; cutoff-value for considering an item answered with guessing probability .25). In addition to that, we examined the corrected item-test-correlations for all response formats and excluded items with $r_{it} < .18$ from the scales (low r_{it}). After that, we computed unidimensional measurement models per subscale and excluded items which considerably deteriorated model fit (*misfit*).

SM Figure 1

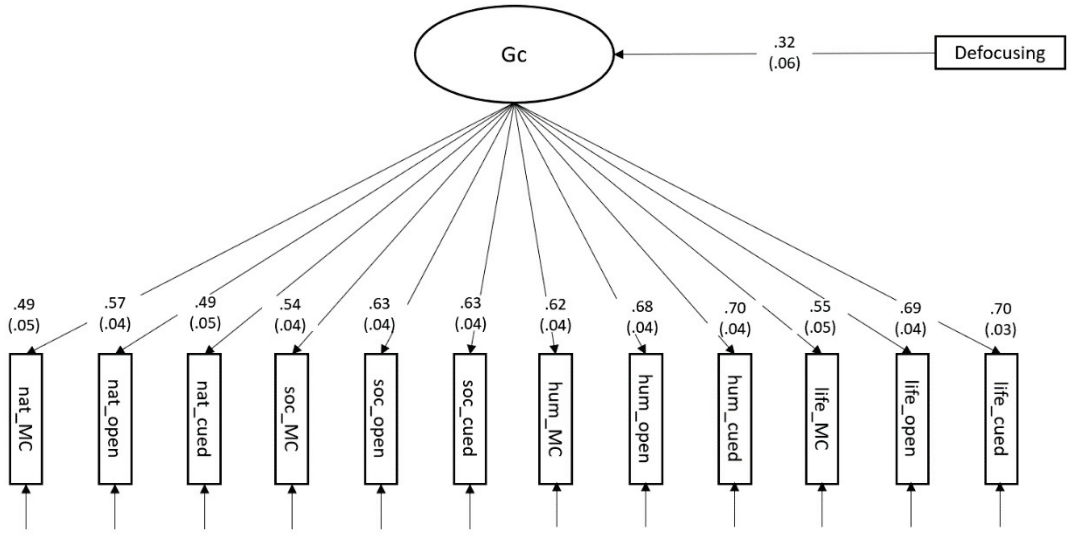
Correlated Factors of the Response Formats with Defocusing.



Note. All parameters are standardized. n = 300; $\chi^2(60)$ = 164.02, CFI = .914, RMSEA = .076, SRMR = .045. Standard errors are depicted in parentheses. Please note that this model is an extension of model A of Table 5 in the manuscript.

SM Figure 2

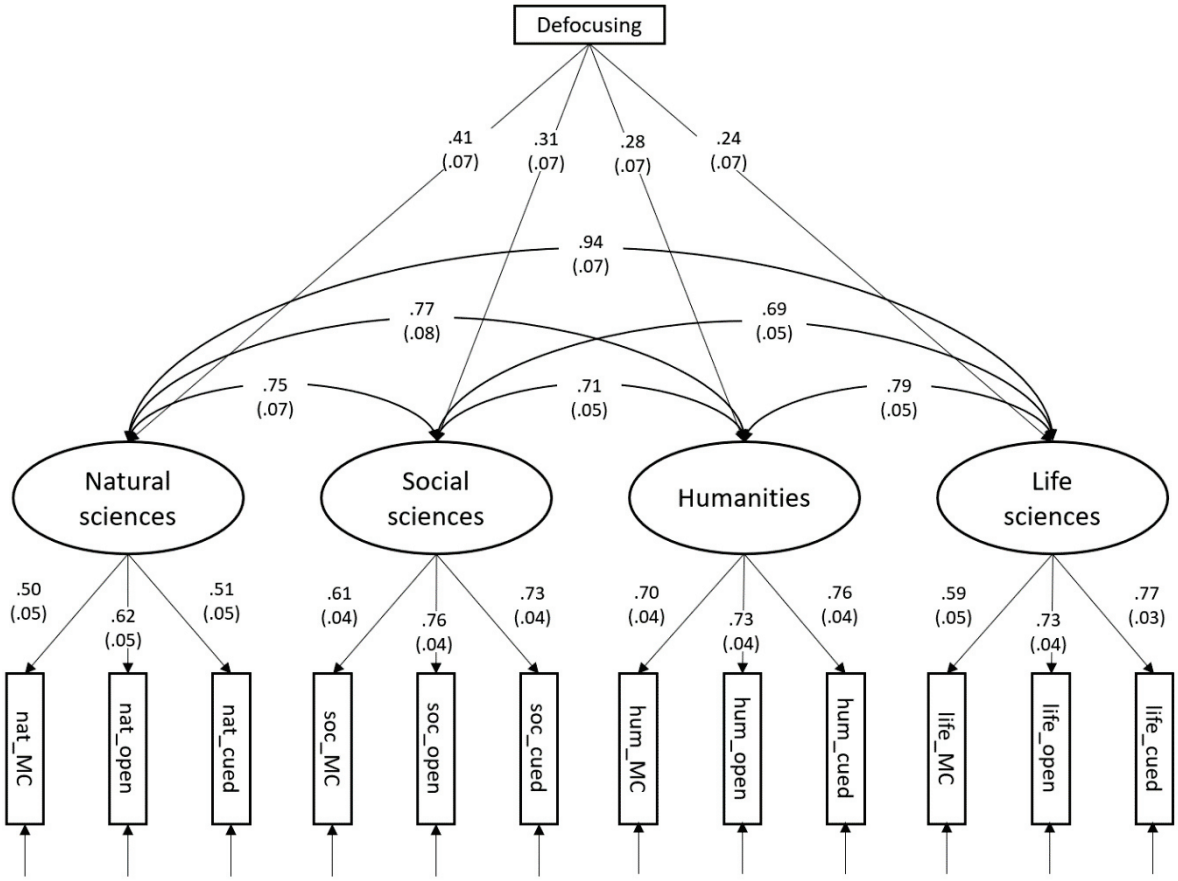
A General Factor Model with Defocusing.



Note. All parameters are standardized. n = 300; $\chi^2(65) = 168.15$, CFI = .914, RMSEA = .073, SRMR = .046. Standard errors are depicted in parentheses. Please note that this model is an extension of model B of Table 5 in the manuscript.

SM Figure 3

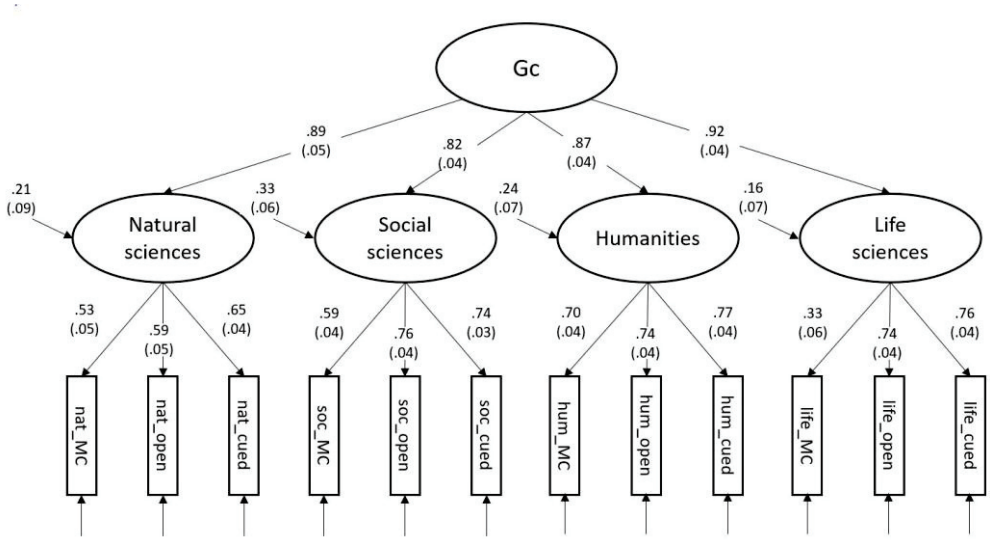
Correlated Factors of the Broad Knowledge Domains Across Response Formats.



Note. All parameters are standardized. n = 300; $\chi^2(56) = 67.04$, CFI = .991, RMSEA = .026, SRMR = .028. Standard errors are depicted in parentheses. Please note that this model is an extension of model C of Table 5 in the manuscript.

SM Figure 4

Higher-Order Model of the Broad Knowledge Domains Across Response Formats.



Note. All parameters are standardized. n = 300; $\chi 2(50) = 62.98$, CFI = .989, RMSEA = .029, SRMR = .029. Standard errors are depicted in parentheses. Please note that this model a depiction of model D of Table 5 in the manuscript. Please note that we specified the same model with age as a predictor of Gc, and found that age was not a substantial predictor of Gc ($\beta = .13$, p = .073; *n* = 300; $\chi 2(61) = 112.44$, CFI = .958, RMSEA = .053, SRMR = .042).

References

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

Steger, D., Schroeders, U., & Wilhelm, O. (2019). On the dimensionality of crystallized intelligence: A smartphone-based assessment. *Intelligence*, *72*, 76–85.

Driving Factors of Individual Differences in Broad Retrieval Ability: *Gr* Is More Than the Sum of Its Parts

Supplementary Material

Method

*Typing Speed***.** The items for assessing typing speed were as follows in German [in English]:

1. Heute hat es viel geregnet. [Today it rained a lot.]
2. Morgen gehe ich nach Hause. [Tomorrow I will go home.]
3. Ich spreche mehrere Sprachen. [I speak several languages].
4. Vatis Schoßhund jagt zwölf Boxkämpfer quer über Sylt. [Dad's lapdog chases twelve boxers across Sylt.]

SM Table 1

Full Manifest Correlation Matrix of All Indicators with Means and Standard Deviations.

	Variable	*Mean*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	Things_l1	70.40	20.84																		
2	Things_l2	72.03	24.95	.75																	
3	Things_s1	11.26	4.06	.56	.58																
4	Things_s2	12.58	5.49	.68	.61	.60															
5	Things_s3	22.20	7.47	.56	.53	.55	.48														
6	Words_l1	75.74	21.65	.65	.69	.53	.49	.49													
7	Words_l2	71.03	28.04	.54	.59	.40	.44	.34	.68												
8	Words_s1	19.53	5.52	.47	.50	.39	.42	.46	.67	.55											
9	Words_s2	13.77	5.18	.44	.48	.31	.31	.31	.56	.55	.51										
10	Words_s3	16.76	5.69	.43	.41	.31	.29	.37	.58	.45	.55	.49									
11	Syno_l1	6.08	3.04	.44	.42	.34	.34	.28	.44	.44	.32	.41	.24								
12	Syno_s1	6.51	2.65	.47	.47	.41	.43	.35	.48	.34	.31	.37	.26	.47							
13	Syno_s2	6.72	3.10	.43	.47	.34	.41	.29	.44	.41	.33	.35	.26	.45	.61						
14	BIS_Item1	11.34	7.13	.30	.34	.30	.25	.23	.33	.26	.23	.18	.22	.16	.28	.25					
15	BIS_Item2	16.58	7.50	.23	.30	.26	.28	.26	.31	.27	.24	.16	.17	.16	.19	.23	.29				
16	BIS_Item3	8.35	5.32	.34	.38	.32	.31	.26	.36	.27	.20	.17	.16	.25	.28	.31	.37	.34			
17	BIS_Item4	12.86	5.26	.39	.45	.36	.37	.32	.41	.32	.32	.24	.23	.24	.33	.36	.45	.39	.41		
18	WMC updating fig	0.75	0.20	.28	.20	.29	.28	.28	.28	.17	.24	.19	.22	.19	.13	.12	.14	.10	.21	.21	
19	WMC updating num	0.69	0.18	.38	.28	.34	.32	.34	.40	.27	.34	.25	.31	.26	.27	.26	.18	.14	.27	.18	.60
20	WMC binding	0.81	0.13	.38	.27	.28	.27	.41	.33	.19	.27	.24	.26	.17	.18	.11	.15	.15	.24	.20	.47
21	SM num1	0.28	0.14	.46	.37	.30	.32	.36	.37	.27	.19	.25	.20	.27	.29	.24	.22	.13	.23	.26	.26
22	SM num2	0.23	0.15	.46	.37	.27	.36	.36	.41	.28	.26	.22	.30	.32	.29	.29	.24	.17	.24	.22	.28
23	SM verb1	0.42	0.25	.48	.45	.31	.41	.33	.43	.35	.25	.27	.26	.34	.35	.36	.23	.19	.31	.33	.34
24	SM verb2	0.38	0.26	.53	.45	.35	.39	.36	.46	.38	.25	.28	.29	.34	.35	.39	.25	.22	.31	.35	.38
25	Comp	1502.26	282.25	-.20	-.13	-.12	-.16	-.17	-.14	-.12	-.15	-.09	-.16	-.09	.01	-.01	-.11	-.14	-.11	-.20	-.13

26	LexDec	650.10	89.65	-.11	-.12	-.11	-.04	-.12	-.13	-.10	-.14	-.10	-.21	-.08	-.10	-.06	-.02	.09	-.04	-.11	.02
27	Search	389.78	46.29	-.11	-.14	-.10	-.10	-.18	-.16	-.06	-.17	-.14	-.19	-.07	-.05	-.07	-.06	-.09	-.04	-.14	-.14
28	Gc.mc_nat	0.59	0.18	.43	.33	.29	.37	.41	.36	.23	.24	.20	.26	.23	.33	.28	.19	.21	.28	.27	.21
29	Gc.mc_soc	0.60	0.14	.25	.17	.18	.24	.26	.17	.10	.12	.14	.12	.19	.16	.12	.05	.08	.11	.10	.17
30	Gc.mc_hum	0.60	0.18	.40	.38	.30	.36	.30	.39	.26	.23	.27	.16	.25	.28	.18	.15	.19	.28	.27	.21
31	Gc.mc_life	0.70	0.19	.33	.33	.37	.44	.30	.28	.25	.23	.18	.17	.26	.28	.24	.11	.16	.17	.19	.14
32	Gc.open_nat	0.49	0.20	.45	.31	.26	.38	.44	.38	.21	.27	.28	.25	.24	.29	.24	.17	.20	.28	.17	.25
33	Gc.open_soc	0.31	0.21	.30	.27	.25	.25	.34	.22	.21	.17	.21	.18	.19	.23	.19	.08	.19	.14	.15	.09
34	Gc.open_hum	0.26	0.18	.46	.35	.28	.40	.37	.38	.29	.26	.26	.24	.32	.28	.25	.17	.18	.22	.20	.16
35	Gc.open_life	0.38	0.21	.43	.43	.37	.46	.34	.39	.29	.32	.33	.23	.33	.34	.31	.20	.17	.24	.24	.23
36	Voc1	0.80	0.15	.45	.37	.35	.40	.33	.40	.29	.28	.30	.32	.25	.34	.27	.28	.10	.24	.22	.21
37	Voc2	0.87	0.12	.39	.31	.38	.37	.24	.30	.24	.22	.23	.20	.22	.23	.24	.22	.10	.15	.18	.25
38	Voc3	0.91	0.11	.45	.41	.40	.39	.38	.40	.25	.26	.28	.29	.28	.34	.29	.25	.18	.33	.31	.25

Note. $p < .05$ for all correlations > .12. All variables refer to the indicators presented in the latent variable models of the original manuscript. WMC bind = Working Memory Cacpacity binding, WMC updt = Working Memory Capacity updating, SM = Secondary Memory, Voc = Vocabulary, Comp = Comparison, LexDec = Lexical Decision, fig = figural, num = numerical, verb = verbal, nat = natural sciences, soc = social sciences, hum = humanities, life = life sciences.

SM Table 1 continued

Full Manifest Correlation Matrix of All Indicators with Means and Standard Deviations.

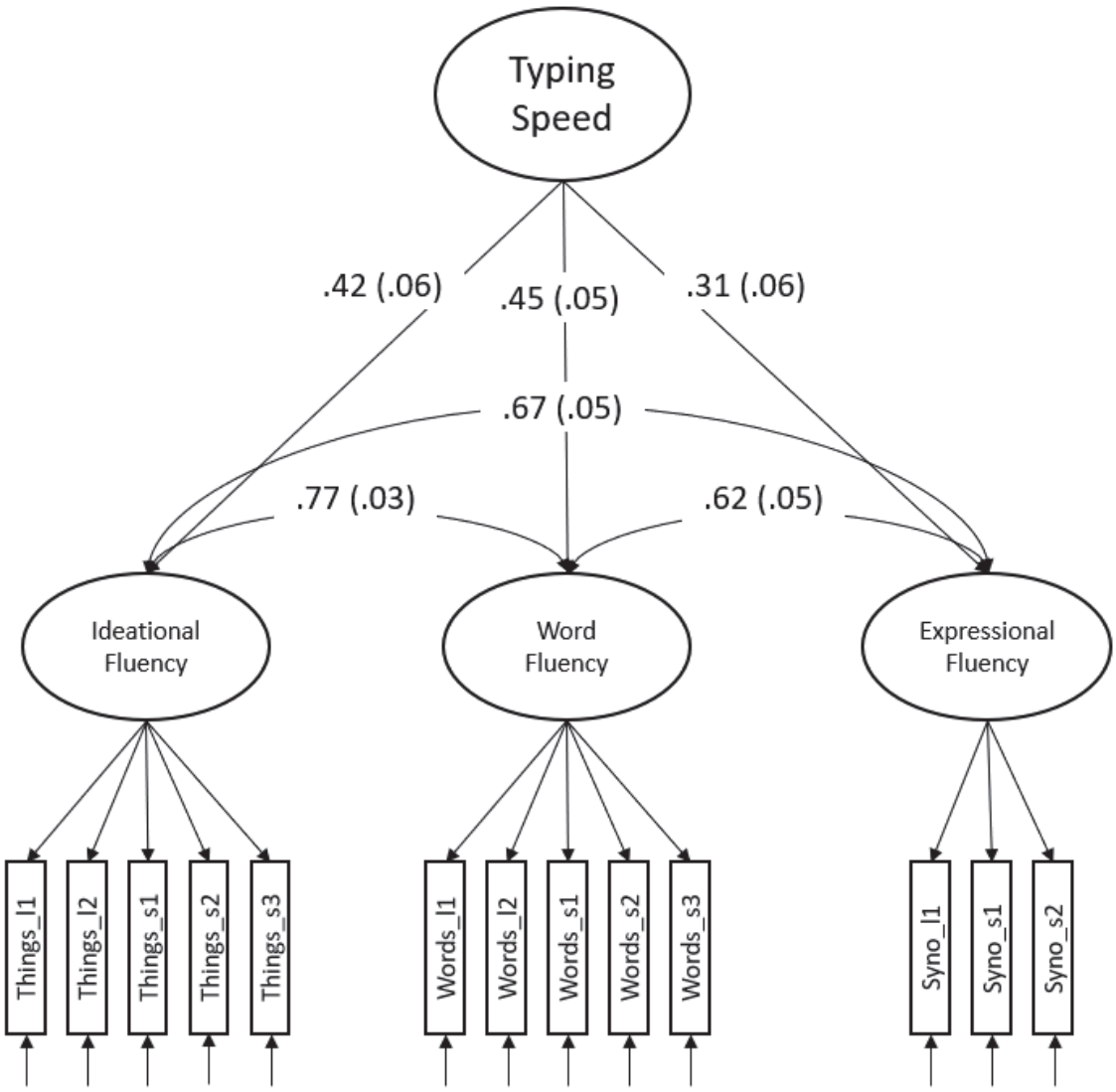
	Variable	*Mean*	*SD*	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
1	Things_l1	70.40	20.84																				
2	Things_l2	72.03	24.95																				
3	Things_s1	11.26	4.06																				
4	Things_s2	12.58	5.49																				
5	Things_s3	22.20	7.47																				
6	Words_l1	75.74	21.65																				
7	Words_l2	71.03	28.04																				
8	Words_s1	19.53	5.52																				
9	Words_s2	13.77	5.18																				
10	Words_s3	16.76	5.69																				
11	Syno_l1	6.08	3.04																				
12	Syno_s1	6.51	2.65																				
13	Syno_s2	6.72	3.10																				
14	BIS_Item1	11.34	7.13																				
15	BIS_Item2	16.58	7.50																				
16	BIS_Item3	8.35	5.32																				
17	BIS_Item4	12.86	5.26																				
18	WMC updating fig	0.75	0.20																				
19	WMC updating num	0.69	0.18																				
20	WMC binding	0.81	0.13	.53																			
21	SM num1	0.28	0.14	.32	.41																		
22	SM num2	0.23	0.15	.37	.34	.61																	
23	SM verb1	0.42	0.25	.39	.35	.65	.66																
24	SM verb2	0.38	0.26	.44	.33	.56	.62	.83															
25	Comp	1502.26	282.25	-.14	-.22	-.11	.02	-.03	-.06														

26	LexDec	650.10	89.65	.08	-.02	.03	.01	.00	-.02	.26											
27	Search	389.78	46.29	-.15	-.21	-.03	-.01	-.07	-.07	.38	.35										
28	Gc.mc_nat	0.59	0.18	.29	.26	.35	.38	.41	.42	-.12	-.04	-.10									
29	Gc.mc_soc	0.60	0.14	.24	.24	.23	.22	.24	.26	-.03	.08	.02	.25								
30	Gc.mc_hum	0.60	0.18	.22	.19	.26	.26	.29	.28	-.05	-.08	.04	.40	.28							
31	Gc.mc_life	0.70	0.19	.23	.16	.21	.24	.28	.27	-.06	-.07	-.02	.40	.24	.35						
32	Gc.open_nat	0.49	0.20	.39	.39	.42	.43	.45	.41	-.15	.02	-.12	.60	.34	.37	.38					
33	Gc.open_soc	0.31	0.21	.19	.18	.22	.19	.24	.26	.04	.03	.05	.37	.41	.38	.32	.38				
34	Gc.open_hum	0.26	0.18	.26	.30	.36	.37	.43	.39	-.05	-.05	-.06	.46	.36	.46	.34	.51	.45			
35	Gc.open_life	0.38	0.21	.26	.25	.31	.34	.40	.36	-.16	-.07	-.09	.44	.27	.42	.59	.46	.34	.37		
36	Voc1	0.80	0.15	.37	.34	.33	.38	.38	.36	.00	-.03	.02	.39	.23	.44	.45	.41	.32	.37	.44	
37	Voc2	0.87	0.12	.31	.21	.21	.19	.27	.29	.05	.01	-.03	.25	.24	.44	.40	.28	.35	.30	.47	.54
38	Voc3	0.91	0.11	.36	.37	.35	.32	.39	.39	-.11	-.08	-.09	.44	.30	.49	.42	.48	.45	.42	.47	.66	.58

Note. $p < .05$ for all correlations > .12. All variables refer to the indicators presented in the latent variable models of the original manuscript. WMC bind = Working Memory Cacpacity binding, WMC updt = Working Memory Capacity updating, SM = Secondary Memory, Voc = Vocabulary, Comp = Comparison, LexDec = Lexical Decision, fig = figural, num = numerical, verb = verbal, nat = natural sciences, soc = social sciences, hum = humanities, life = life sciences.

SM Figure 1

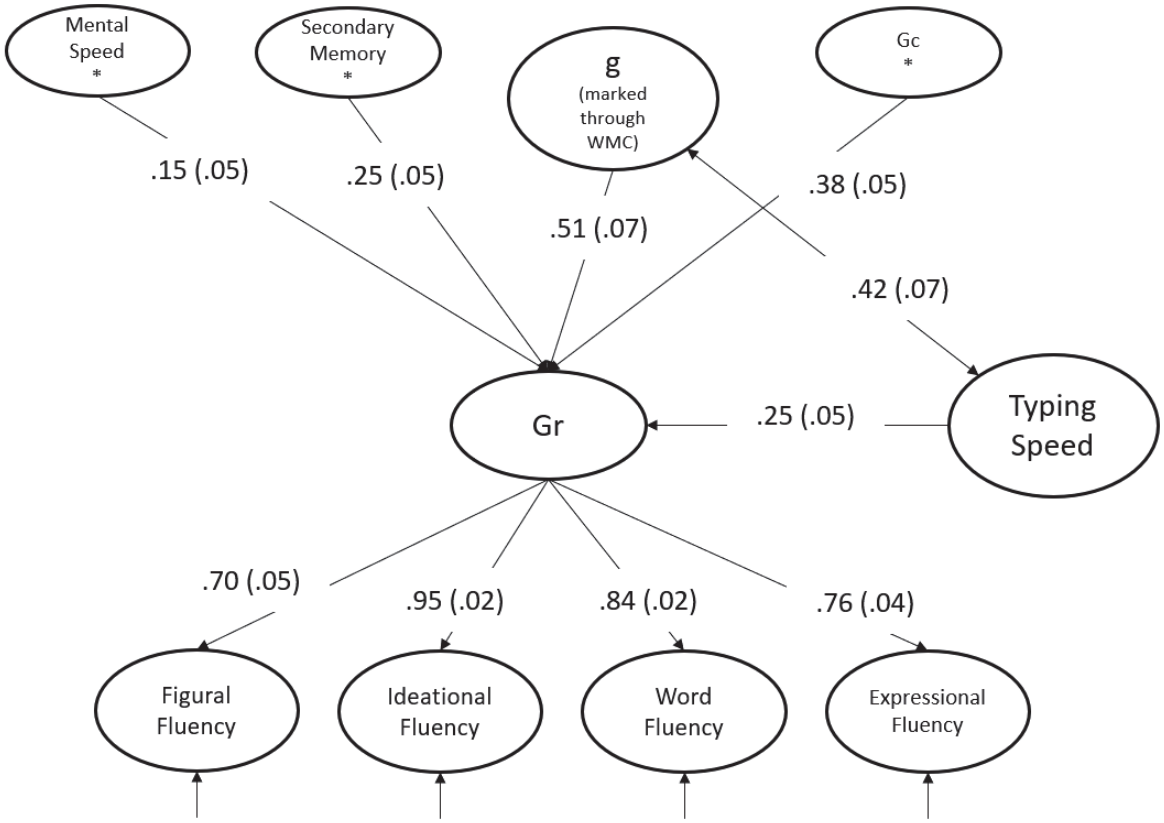
Measurement Model of Gr with Three Correlated Factors and Typing Speed as a Predictor.



Note. All parameters are standardized. n = 322; $\chi^2(113) = 245.88$, CFI = .943, RMSEA = .060, SRMR = .048. Standard errors are depicted in parentheses. Please note that for ease of interpretation the regression weight of the typing speed towards ideational, word, and expressional fluency was multiplied by -1 so that positive loadings reflect better performance (i.e., faster reaction times are better).

SM Figure 2

Latent Regression Model of Gr Typing Speed as Control Variable.



Note. All parameters are standardized. n = 322; $\chi^2(56)$ = 1278.41, CFI = .920, RMSEA = .044, SRMR = .057. Standard errors are depicted in parentheses. Please note that for ease of interpretation the regression weight of the typing speed towards Gr and the correlation between *g* and typing speed was multiplied by -1 so that positive loadings reflect better performance (i.e., faster reaction times are better).