ORIGINAL RESEARCH
published: 22 July 2022
doi: 10.3389/feduc.2022.909818

Development and Initial Validation of an Admission Test for Bachelor Psychology Studies

Luc Watrin[1]*, *Mattis Geiger*[1,2], *Julie Levacher*[3], *Birgit Spinath*[4] *and Oliver Wilhelm*[1]

[1] Department of Individual Differences and Psychological Assessment, University of Ulm, Ulm, Germany, [2] Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany, [3] Department of Individual Differences and Psychodiagnostics, Saarland University, Saarbrücken, Germany, [4] Department of Psychology, Heidelberg University, Heidelberg, Germany

Extensive evidence clearly endorses the use of standardized reasoning ability tests and subject-specific knowledge tests as valid and useful tools for admission-restricted study programs. Yet, tests are still rarely applied for university admission in Germany. Instead, current admission practices are predominantly based on grade point average (GPA) achieved in high school. In the present study, we report the development and validation of a test battery for admission into bachelor's degree programs in psychology for German universities. Its compilation is driven by evidence from international validity generalization, consensual models of cognitive abilities, and a taxonomy of the B.Sc. psychology degree in Germany. It consists of three subtests for reasoning ability, two tests that tap relevant declarative knowledge, and two psychology-specific text comprehension tests. N = 371 freshmen from five German universities completed the tests and university GPA was retrieved 2.5 years later. We use confirmatory factor analyses and structural equation modeling to investigate the construct and criterion validity of the test battery. The results indicate that individual tests, as well as the test battery, meet psychometric requirements. As expected, the test battery predicts university GPA substantially and incrementally beyond high school GPA. The results illustrate the substantial added value that standardized achievement tests provide in university admissions.

Keywords: student selection, cognitive abilities, text comprehension, knowledge, criterion validity

OPEN ACCESS

Edited by:
Johanna Hissbach,
University of Hamburg, Germany

Reviewed by:
Florian Krieger,
Technical University Dortmund,
Germany
Marco Koch,
University of Saarland, Germany
Dietrich Klusmann,
University Medical Center
Hamburg-Eppendorf, Germany

****Correspondence:***
Luc Watrin
luc.watrin@uni-ulm.de

Specialty section:
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

Received: *31 March 2022*
Accepted: *17 June 2022*
Published: *22 July 2022*

Citation:
Watrin L, Geiger M, Levacher J,
Spinath B and Wilhelm O (2022)
Development and Initial Validation
of an Admission Test for Bachelor
Psychology Studies.
Front. Educ. 7:909818.
doi: 10.3389/feduc.2022.909818

INTRODUCTION

From an egalitarian perspective, university admission should be based on random selection to offer all applicants the same chance of obtaining a university place. This perspective, however, neglects evident individual differences in the ability to successfully obtain an academic degree. Individual differences in mastering the contents of an academic degree are of great importance to both the individual and the society. Therefore, university admission (in Germany) is geared toward a largely meritocratic system (Perfetto et al., 1999)—the admission process prioritizes those individuals whose expected aptitude to successfully complete the studies is considered the highest. Admission to capacity-restricted degrees in Germany is primarily based on success in high school, a procedure that is legitimized by comprehensive evidence that the grade point (GPA) in high school is a strong predictor of university success (e.g., Trapmann et al., 2007). School grades are highly valid and readily available, but they also suffer from a lack of comparability across schools and states (Köller et al., 2004). Additionally, they are correlated with sex (Voyer and Voyer, 2014) and socioeconomic status (Autorengruppe Bildungsberichterstattung, 2020). As there is no convincing evidence for a causal effect of sex or parent's SES on school (or university) achievement, group differences in school grades with respect to these variables are problematic because they are the basis for top-down admission to university. Therefore, alternative or supplementary selection procedures have been proposed, of which subject-specific aptitude tests have proven to be the most valid and useful (Camara and Kimmel(eds), 2005).

Despite the overwhelming scientific evidence, aptitude tests have not made their way into the highly selective bachelor psychology studies in Germany, but jurisdiction and policymakers have recently demanded to improve admission processes for restricted studies beyond GPA (Bundesverfassungsgericht [BVerfG], 2017). Therefore, we report the development and initial validation of an admission test for bachelor psychology students. Thereto, we build on the extensive literature on subject-specific aptitude tests and contemporary models of cognitive ability. In the following, we first discuss general requirements for admission procedures and which types of measures are suitable. Subsequently, we discuss models of cognitive ability and how they relate to aptitude tests prevalent in university admission. Finally, we describe our rationale for the development of a psychology-specific aptitude test.

Procedures used to make admission (i.e., selection) decisions must meet strict psychometric and legal requirements because of their far-reaching consequences (e.g., access to professions, salary, social status, but also potential lawsuits). These requirements have been described in several professional guidelines and norms (e.g., International Organization for Standardization [ISO], 2011; American Educational Research Association [AERA] et al., 2014; DIN, 2016; Tippins et al., 2018). In there, the focus lies on different aspects of the admission process such as the characteristics of the procedure itself (e.g., objectivity, reliability, validity, fairness) and qualifications of the responsible persons (e.g., familiarity with standardized behavioral observation or psychometric testing). Although choosing an appropriate method to select students might initially seem complicated, the boundary conditions of university admission processes and the extant scientific literature on the utility of different methods simplify the choice considerably.

First, an admission process must be insusceptible to intentional distortions to ensure that applicants are assessed appropriately and fairly. Because admission testing is a high-stakes situation, it is in the interest of all applicants to present themselves in a way that maximizes their chances of admission. This potentially includes purposefully embellishing or misstating one's accomplishments, attributes, and abilities. Consequently, all methods based on self-report data that cannot be verified are unsuitable. This includes all measures of *typical performance* (Cronbach, 1960), such as questionnaires of personality or interests. While some traits typically measured in self-report are certainly relevant for academic and career success, extensive research has shown that self-report measures are susceptible to faking (Viswesvaran and Ones, 1999), that applicants do fake in high-stakes situations, and that this affects selection decisions (Donovan et al., 2014). Principally suited, on the other hand, is information that can be verified (e.g., grades, work experience), as well as measurements of *maximum performance*. In the latter, applicants are confronted with different tasks (e.g., intelligence tests, interviews, assessment centers) and instructed to do as well as they can. In proctored testing, it is therefore impossible for applicants to falsify test results upwards in their favor. In turn, prior test training and coaching can lead to increases in test scores (e.g., Kulik et al., 1984; Levacher et al., 2022) but do not improve cognitive ability (e.g., Estrada et al., 2015). It is therefore essential to give all applicants the same opportunity to comprehensively familiarize themselves with the tests free of charge in order to allow for a fair competition (American Educational Research Association [AERA] et al., 2014).

Second, admission procedures for selective study programs must be scalable to large numbers of applicants. Given the financial and personnel resources of universities, this reduces the choice of admission procedures to those that entail justifiable financial, time, and personnel expenses. Work samples and standardized interviews, for example, have proven to predict job performance (e.g., McDaniel et al., 1994) and to some degree academic achievement (e.g., Hell et al., 2007b). However, the resource requirements of such interactive methods are arguably prohibitive in the case of bachelor psychology student admission in Germany—in the state of Baden-Württemberg, for example, universities face approximately 17,000 applications for 700 study places. For smaller admission problems, for example in specialized master's programs or Ph.D. programs, interactive procedures may be justifiable and useful, but in mass processes such as the admission to bachelor's programs, they are hardly justifiable from a cost-benefit perspective.

Finally, we argue that predictive validity is key in any evaluation of admission procedures. Because the degree of success in higher education is essential for both the individual and the institutions, the utility of admission procedures should mainly be evaluated by their ability to predict academic achievement—above and beyond high-school GPA, which is already routinely used and which is an established predictor of academic achievement (Trapmann et al., 2007).

If the requirements of incremental predictive validity, efficiency, and robustness against intentional distortions are considered jointly, the extant scientific literature clearly speaks for the use of standardized ability tests (Kuncel et al., 2001; Westrick et al., 2015; Beard and Marini, 2018).

Research on individual differences in cognitive abilities has developed consensual theoretical models that are widely accepted (e.g., Carroll, 1993; Schneider and McGrew, 2018). Therein, intelligence is conceptualized as a hierarchical construct with a strong general factor of intelligence (*g*) at the top that explains positive correlations amongst lower-order cognitive factors. Of the more specific cognitive abilities below the apex, fluid (*gf*) and crystallized intelligence (*gc*) are particularly relevant. *Gf* represents the decontextualized ability to solve abstract problems and is elementary for knowledge acquisition of any kind (Cattell, 1987; Wilhelm and Kyllonen, 2021). It is best measured with reasoning tasks (Wilhelm, 2005b) and has been equated with the general factor of intelligence for psychometric reasons (e.g., Gustafsson, 1984). *Gc*, in turn, reflects acquired skills and knowledge in different domains (e.g., Cattell, 1987) and is best measured with declarative knowledge tests (Schipolowski et al., 2014). While knowledge acquisition requires the investment of fluid abilities to some extent (Cattell, 1987), it also has additional predictors (interests, personality traits, and learning opportunities; Ackerman, 1996). Thus, *gc* has the potential to contribute to the prediction of criteria such as academic achievement above and beyond *gf*. For example, Postlethwaite (2011) found stronger associations of *gc* and academic achievement than for *gf*.

In university admission, cognitive ability tests often go by the name of *aptitude tests*, which aim to measure the intellectual abilities necessary to successfully complete studies. For example, the *Scholastic Assessment Test* (SAT), the *American College Test* (ACT), and the *Graduate Record Examination* (GRE) are institutionalized in college and university admission in the United States. The term aptitude test rather stems from an unfortunate disconnection between educational and psychometric research than a real conceptual or empirical difference to psychometric intelligence tests. From an empirical perspective, aptitude tests are hardly distinguishable from conventional cognitive ability tests (e.g., Coyle, 2006; Koenig et al., 2008). From a theoretical perspective, any aptitude test can be described in terms of established ability factors, or a linear combination thereof. Thereto, it is helpful to locate the tests on a theoretical continuum from decontextualized to contextualized abilities, depending on how much they strain reasoning and factual knowledge. *Gf* is close to the decontextualized end of the continuum and fact knowledge tests, as indicators of *gc*, represent the contextualized end. For example, "general" aptitude tests such as the SAT-I and the GRE General primarily measure verbal and quantitative reasoning (i.e., *gf* with aspects of word or number knowledge). Common measures of reading comprehension, in turn, rely on both *gf* and *gc* (Schroeders and Wilhelm, 2012). "Subject-specific" aptitude tests, such as the SAT-II and the GRE Subject Test are tests that predominantly measure knowledge in a particular field (e.g., chemistry, physics) and are best understood as measures of *gc*.

There is abundant empirical evidence that such ability tests predict academic achievement (as measured by university grades) above and beyond high school GPA (Bridgeman et al., 2000; Kuncel et al., 2001; Westrick et al., 2015; Beard and Marini, 2018). Due to their general applicability across disciplines and due to their long-standing institutionalization, there is more evidence for tests such as the SAT-I or the GRE General. Subject-specific tests like the SAT-II or the GRE Subject Tests are also providing incremental value in predicting academic achievement (e.g., Kuncel et al., 2001).

Empirical evidence clearly speaks for the generalizability of these findings across countries and disciplines, but with regard to the present undertaking we will briefly narrow our focus on findings from German-speaking countries and psychology in particular. Except for admission to medical degrees, aptitude tests are not common in German-speaking countries (Austria, Germany, Switzerland). Still, two meta-analyses for these countries are generally in line with results from international research and underline the predictive validity of subject-specific admission tests (Hell et al., 2007a; Schult et al., 2019). The most notable test development for psychology-specific admissions in Germany has been reported by Formazin et al. [2011; but see also Schmidt-Atzert (2005) and Heene (2007)]. The authors developed a comprehensive test battery of reasoning, knowledge, and text comprehension tests for which they established a measurement model of cognitive abilities with a strong general factor and a nested factor of crystallized intelligence. Critically, both factors predicted academic achievement ($\beta_G = 0.32$, $\beta_{Gc} = 0.59$, $R^2_{Total} = 44\%$).

In sum, empirical evidence clearly demonstrates the utility of standardized ability tests in university admission, and psychology studies in Germany more specifically. Importantly, key characteristics of an admission test can be deduced from extant findings. Given its fundamental importance for learning, fluid intelligence is relevant in any study program and is reliably measurable with reasoning tests. To increase acceptance (Kersting, 2008), the typically decontextualized reasoning test can be contextualized by adapting superficial characteristics. Declarative knowledge tests are the means of choice to test aspects of crystallized abilities that are deemed relevant for a degree (Cattell, 1987; Schipolowski et al., 2014). The development of knowledge tests needs to be subject-specific and the selection of knowledge domains and subdomains is best based on a requirement analysis to maximize content and potentially criterion validity. Finally, text comprehension tests are an attractive addition to pure measures of *gf* and *gc* because they require both abilities (Schroeders and Wilhelm, 2012), possess high face validity, and have a track record of successful application in university admission. Considering their *gc* share, subject-specific test development is advised. To maximize fit with study content, it is recommended to develop a taxonomy of study content, which then informs the text content of the test.

In the present study, we compile, develop, and validate a psychology-specific admission test. This test consists of several reasoning, comprehension, and knowledge tests which are contextualized for the psychology studies to varying degrees and thus builds on the work of Formazin et al. (2011). We provide evidence for its reliability, construct validity, and, importantly, its criterion validity. Thereto, we use university GPA 2.5 years later.

MATERIALS AND METHODS

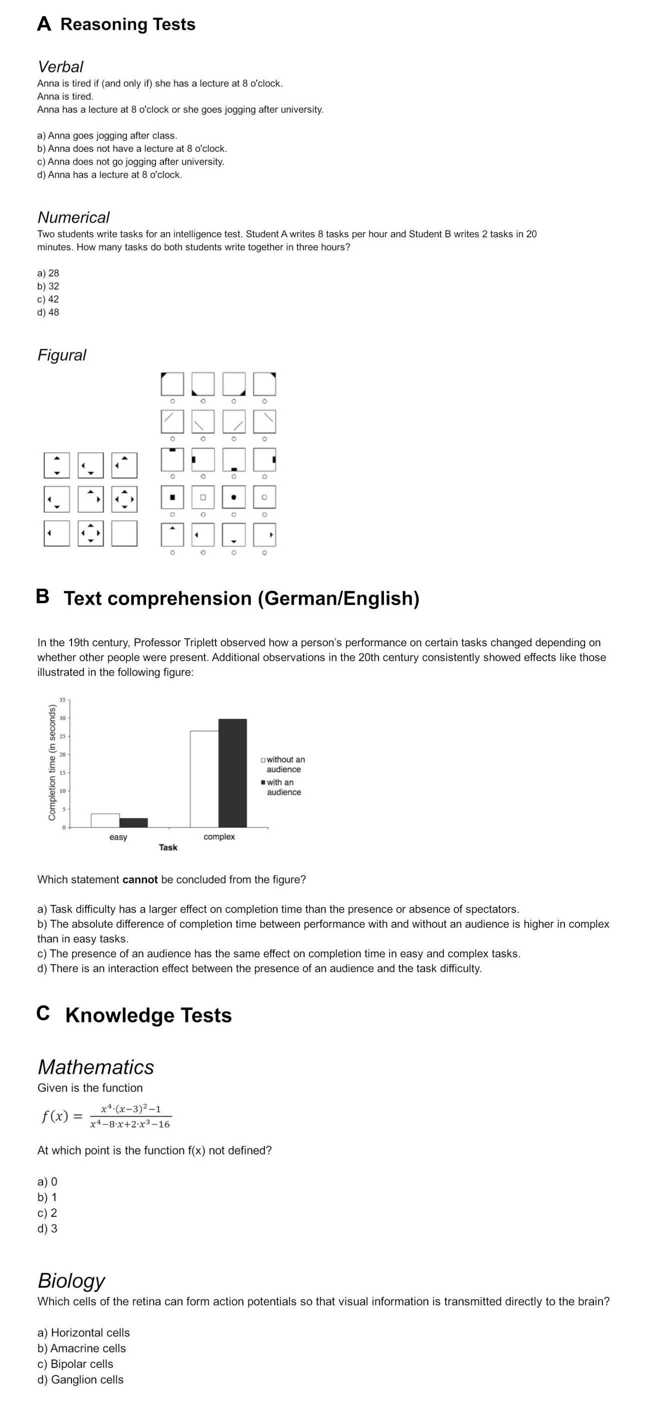
Sample

The study was reviewed and approved by the local ethics board (Spin 2019 2/1). Participants were $N = 371$ psychology freshmen from five universities in south-western Germany. As expected in a sample of German first-year psychology students, participants were young ($M_{age} = 21.3$, $SD_{age} = 3.5$) and mostly female (84%). Demographic variables and sample sizes by university are reported in **Table 1**.

TABLE 1 | Demographic variables and sample sizes for tests, high school grade point average (GPA), and University GPA by University.

	University				
	A	B	C	D	E
Gender [m/f/other]	14/46/0[a]	10/48/0	7/53/0	8/50/0	21/111/2
Age [*M* (*SD*)]	22.1 (4.8)	20.9 (3.0)	21.7 (4.2)	20.7 (1.7)	21.3 (3.2)
N_{Tests}	61	58	60	58	134
$N_{\text{Highschool GPA}}$	60	56	59	57	131
$N_{\text{University GPA}}$	48	37	31	33	65

[a] *Information about the gender of one person is missing.*



A Reasoning Tests

Verbal

Anna is tired if (and only if) she has a lecture at 8 o'clock.
Anna is tired.
Anna has a lecture at 8 o'clock or she goes jogging after university.

a) Anna goes jogging after class.
b) Anna does not have a lecture at 8 o'clock.
c) Anna does not go jogging after university.
d) Anna has a lecture at 8 o'clock.

Numerical

Two students write tasks for an intelligence test. Student A writes 8 tasks per hour and Student B writes 2 tasks in 20 minutes. How many tasks do both students write together in three hours?

a) 28
b) 32
c) 42
d) 48

Figural

B Text comprehension (German/English)

In the 19th century, Professor Triplett observed how a person's performance on certain tasks changed depending on whether other people were present. Additional observations in the 20th century consistently showed effects like those illustrated in the following figure:

Which statement **cannot** be concluded from the figure?

a) Task difficulty has a larger effect on completion time than the presence or absence of spectators.
b) The absolute difference of completion time between performance with and without an audience is higher in complex than in easy tasks.
c) The presence of an audience has the same effect on completion time in easy and complex tasks.
d) There is an interaction effect between the presence of an audience and the task difficulty.

C Knowledge Tests

Mathematics

Given is the function

$$f(x) = \frac{x^4 \cdot (x-3)^2 - 1}{x^4 - 8 \cdot x + 2 \cdot x^3 - 16}$$

At which point is the function f(x) not defined?

a) 0
b) 1
c) 2
d) 3

Biology

Which cells of the retina can form action potentials so that visual information is transmitted directly to the brain?

a) Horizontal cells
b) Amacrine cells
c) Bipolar cells
d) Ganglion cells

FIGURE 1 | Sample items of **(A)** verbal, numerical, and figural reasoning tests, **(B)** psychology-specific text comprehension tests, and **(C)** declarative knowledge tests.

Test Battery

Reasoning Tasks

For the reasoning tests, we followed recommendations from the Berlin Intelligence Structure model (Jäger, 1982) to select multiple tests with varying stimulus modalities. Thus, we chose three different *gf* tests with figural, numerical, and verbal content (Wilhelm, 2005b; see **Figure 1A**). The 15-item numerical reasoning test consists of arithmetic text problems that must be solved without aids (e.g., notes, calculator). In the 15-item verbal reasoning test, the correct conclusion must be drawn from a set of premises. To increase the face validity, and thus potentially the acceptance of the tests (Kersting, 2008), we superficially contextualized the verbal and numerical tasks for psychology or the general university context where possible. The figural reasoning task comprises 28 figural matrices in which rules must be recognized according to which figures systematically change within a 3 × 3 matrix (Becker et al., 2016). Unlike usual matrices tasks, the answer must be constructed by the user instead of choosing from a fixed set of response options.

Psychology-Specific Text Comprehension

We developed two psychology-specific text comprehension tests in German and English language. We chose English as a second language because it has been identified as a relevant skill in a requirement analysis for psychology studies (Wetzenstein, 2004). This is because it is the predominant language in science and thus the majority of literature that is dealt with during psychology studies in Germany.

To identify relevant content areas for testlets, we first developed a comprehensive taxonomy of the Bachelor Psychology curriculum. The taxonomy was based on the specifications of the German Society for Psychology (DGPs), the association of psychologists active in research and teaching, which defined central subjects of the Bachelor Psychology program (Deutschen Gesellschaft für Psychologie [DGPs], 2022). According to the DPGs guidelines, the subjects General Psychology, Developmental Psychology, Biological Psychology, Methodology, Differential/Personality Psychology, Social Psychology, Psychological Assessment, Industrial and Organizational Psychology, Clinical Psychology, and Educational Psychology must be taught as part of the bachelor psychology curriculum. Based on this list of subjects, the first two authors performed a comprehensive search in textbooks and university calendars to identify the major topics and subtopics within each subject. To validate the taxonomy developed in this way, all professors of psychology working at German universities ($N = 557$) were invited to evaluate the completeness and relevance of the identified topics in an online survey, of which 68 (12.2%) responded (results of this survey are reported at https://osf.io/n9qt8/). Based on these responses, the taxonomy was revised and finalized. For the text comprehension tests, the content of the item stems was selected based on the relevance rating from the survey. Both the German and the English test contain five testlets with three multiple-choice items each (**Figure 1B**).

Study-Specific Prior Knowledge

Based on a requirement analysis (Wetzenstein, 2004) and an empirical investigation of knowledge tests in psychology admission (Kunina et al., 2007) we included declarative knowledge tests of high school level mathematics and biology in the test battery (**Figure 1C**). For the biology test, topics were selected that are covered as part of the bachelor's curriculum in psychology (i.e., "Anatomy," "Evolution," "Reproduction and Development," "Genetics," "Neurobiology," "Behavioral Biology," "Metabolism," and "Cell Biology."). The same approach was chosen for the mathematics test (i.e., "Algebra," "Analysis," "Analytic Geometry," and "Stochastics").

Study Design

At each university, participants were tested in a single session and on a single day within the first 2 weeks of the semester. Test participation was voluntary and compensated (course credits or 40€). Each session was led by a trained test administrator and several subordinate test supervisors to ensure test security and to mimic real testing conditions as closely as possible. Because all but one university imposed pragmatic time constraints of 3 h for the study, we implemented a planned-missingness design (see **Table 2**). Subsets of tests were administered at four universities and all tests were administered at one university. Tests were compiled to estimate all covariances between tests with sufficient power.

Analysis

All analyses were performed in R [version 4.0.4, R Core Team (2020)]. General data handling and visualization were performed with packages from the *tidyverse* [version 1.3.1, Wickham et al. (2019)], descriptive statistics and basic psychometrics were computed with the package *psych* [version 2.1.6, Revelle (2020)] and latent factor models were estimated with the package *lavaan* [version 0.6-9; Rosseel (2012)]. All items were scored 1 if the response was correct, 0 if the item was presented and the answer was false or not answered, and missing in all other cases. Total scores for the subtests were computed as the mean scores of the respective items and the total score of the test battery was computed by averaging all available subtest scores. High-school GPA was scaled across the entire sample to reflect its quality as uniform university entrance qualification. University GPA was scaled within universities. Latent factor models with dichotomous indicators were estimated with *Diagonally Weighted Least Squares* (DWLS) estimation and missing data was handled pairwise. The analysis of models with continuous indicators was carried out using *Full Information Maximum Likelihood* (FIML) estimation to handle data missing by design.

RESULTS

Descriptive Results and Construct Validity

Descriptive statistics of all tests are reported in **Table 3**. As will be the case in future "real" applications, participants completed the core set of items (e.g., 15 items of verbal reasoning) plus an additional third of items to be validated for future test renewal (e.g., five additional items of verbal reasoning). The present results refer to the core itemset. Due to the planned-missingness design, sample sizes for the individual tests ranged between $N = 247$ (knowledge tests) and $N = 357$ (figural reasoning).

Concerning test difficulty, please note that the current sample was highly preselected (based on high school GPA) and thus presumably more cognitively capable than the prospective, less selected, applicant pool. Additionally, whereas test development usually aims to optimize tests for average ability levels, the tests developed here need to differentiate well amongst the upper 10–20% of applicants. With this in mind, the distributions of test scores were generally adequate. Neither the most difficult test (numerical reasoning), nor the easiest test (text comprehension, German) exhibited severe floor or ceiling effects (**Table 3**). The proportion of missing responses per test ranged between 0.6% and 15.8%, indicating that test time provided was sufficient overall. Average item-total correlations range between 0.40 (numerical reasoning, biology knowledge) and 0.74 (figural reasoning), which can be considered good.

Unidimensional measurement models were fitted for all tests. Model fit was considered good with CFI $\geq$ 0.96 and RMSEA $\leq$ 0.06 (Yu, 2002). Model fit ranged from good to excellent, except for the biology test where the CFI was below the suggested cut-off value. Because the misfit was mainly caused by a single residual correlation, we refrained from altering the test. McDonald's (1999) omega was above 0.70 for all tests, indicating good saturation. In sum, unidimensionality and reliability can be assumed for all.

Zero-order correlations of all tests and criteria are reported in **Table 4**. As expected, all tests were moderately positively correlated. Conclusively, the highest correlation was observed between numerical reasoning and the mathematics knowledge test, both rely on mathematic operations, and the two text comprehension tests. The biology test, as a pure measure of fact knowledge, correlated the weakest with the other tests, which all have a strong share of *gf*.

To investigate the structure of the test battery, we compared a nested factor model as established in Formazin et al. (2011) with a more parsimonious *g* factor model. In the nested factor model, all seven tasks loaded on a common *g* factor, and the text comprehension and knowledge tasks additionally loaded on an orthogonal *gc* factor. In this model, the *gc* factor captures the common variance of its tasks that is not captured by the general factor. We used *full maximum likelihood* estimation (Graham, 2009) to account for planned-missingness. The nested

TABLE 2 | Planned-missingness design of the study.

University	Reasoning			Psychology-specific text comprehension		Knowledge	
	Numerical	Verbal	Figural	German	English	Mathematics	Biology
A		✓	✓	✓			✓
B		✓	✓		✓	✓	
C	✓		✓		✓	✓	✓
D	✓	✓	✓	✓			
E	✓	✓	✓	✓	✓	✓	✓

Ticks indicate that a test was administered.

TABLE 3 | Descriptive statistics and fit indices of confirmatory factor analysis.

	Reasoning			Text comprehension		Knowledge	
	Numerical	Verbal	Figural	German	English	Mathematics	Biology
N	250	308	357	250	251	247	247
# Items	15	15	28	15	15	15	15
M (SD) score	6.95 (2.70)	10.86 (2.72)	19.75 (5.88)	12.08 (2.19)	11.18 (2.54)	8.97 (3.01)	9.66 (2.53)
M (SD) r_{it}	0.40 (0.13)	0.46 (0.11)	0.74 (0.12)	0.49 (0.19)	0.51 (0.12)	0.47 (0.13)	0.40 (0.15)
χ^2 of *g*-model	95.81	97.65	720.72	94.89	91.00	79.56	109.86
df of *g*-model	90	90	350	90	90	90	90
p	0.318	0.273	0.000	0.342	0.451	0.777	0.076
CFI	0.971	0.977	0.987	0.987	0.998	1.000	0.912
RMSEA	0.016	0.017	0.054	0.015	0.007	0.000	0.030
ω_{total} factor	0.72	0.80	0.98	0.83	0.84	0.78	0.74

r_{it}, corrected item-total correlation; df, degrees of freedom; CFI, comparative fit index; RMSEA, root mean square error of approximation; ω, McDonald's (1999) omega total reliability coefficient.

TABLE 4 | Zero-order correlations of all tests and criteria.

Variable	1	2	3	4	5	6	7	8	9	10
1. Reasoning—Numerical	250	190	238	190	192	188	187	250	245	128
2. Reasoning—Verbal	0.38	308	295	250	191	187	188	308	301	181
3. Reasoning—Figural	0.24	0.41	357	238	238	238	239	357	349	209
4. Text comprehension—German	0.36	0.37	0.38	250	133	129	188	250	245	144
5. Text comprehension—English	0.37	0.31	0.31	0.48	251	247	188	251	245	133
6. Knowledge—Mathematics	0.46	0.35	0.22	0.40	0.29	247	188	247	241	130
7. Knowledge—Biology	0.08^t	0.02^t	0.12^t	0.17	0.21	0.15	247	247	242	139
8. Total test score	0.69	0.70	0.66	0.72	0.71	0.66	0.46	370	362	213
9. GPA—High school	0.16	0.19	0.17	0.24	0.11^t	0.11^t	0.08^t	0.21	363	209
10. GPA—University	0.09^t	0.21	0.31	0.35	0.36	0.19	0.22	0.38	0.35	214

Values below the diagonal are bivariate correlations. Values on and above the diagonal are the respective sample size for the correlations. Grades were recoded so that higher values indicate better performance. $^t p > 0.05$.

factor model fit the data adequately ($\chi^2_{[10]}$ = 19.88, p = 0.030, CFI = 0.966, RMSEA = 0.052). Although the factor loadings were comparable to those reported by Formazin et al. (2011; $\Delta\lambda_{text\_de}$ = –0.06, $\Delta\lambda_{math}$ = –0.23, $\Delta\lambda_{text\_en}$ = –0.03, $\Delta\lambda_{bio}$ = –0.06), the factor variance was not significant. Thus, a reliable *gc* factor could not be established next to the strong general factor in the current data. A power analysis using Monte-Carlo simulation indicated that this was likely due to the comparatively small sample size of the current study. A sample size of approximately 1300 [cf. a sample of N = 1187 in Formazin et al. (2011)] would have been necessary to achieve a power of 0.80 for establishing a factor variance larger than zero.[1]

The more parsimonious *g* factor model had a fit of $\chi^2_{(14)}$ = 28.88, p = 0.011, CFI = 0.949, RMSEA = 0.054. As conceivable from the manifest correlations, modification indices pointed toward a residual correlation between numerical

[1] https://osf.io/n9qt8/

reasoning and mathematics as the main source of the misfit. Allowing for this sensible correlation representing the indicators' shared numerical content improved the fit significantly: $\chi^2_{(13)} = 20.62$, $p = 0.081$, CFI = 0.974, RMSEA = 0.040. Because it was adequate for the available sample size and explained the observed data well, we proceeded with this measurement model for the investigation of criterion validity.

Predictive Validity

Manifest correlations indicated significant associations between the total test score, high-school GPA and university GPA. The total test score correlated $r = 0.21$ ($p < 0.001$) with high school GPA and, more importantly, $r = 0.38$ ($p < 0.001$) with university GPA after 2.5 years. The correlation of high school GPA and university GPA was in the same range ($r = 0.35$, $p < 0.001$).

To investigate incremental validity, we estimated latent factor models (see the **Appendix** for manifest simple and multiple regression results). In a first model (Model A, **Figure 2A**), we tested the incremental validity of *g* above and beyond high school GPA, using the method proposed by Feng and Hancock (2021). Thereto, we regressed *g* on high school GPA and create latent phantom variables (ξ_{School}, ξ_g) which capture the variance of high school GPA and the residual variance of *g* after controlling for high school GPA, respectively. These two variables were regressed on university GPA as the criterion [see Feng and Hancock (2021) for details on the parameterization of such models]. This model fit the data well: $\chi^2_{(25)} = 40.68$, $p = 0.025$, CFI = 0.958, RMSEA = 0.041. A *post-hoc* power analysis indicated that the sample size was sufficient to estimate all parameters of interest. The implemented modeling has the advantage that path coefficients can readily be interpreted in terms of explained variance and incrementally explained variance: High school GPA predicted $R^2 = 0.382^2 = 14.7\%$ ($p = 0.001$) of the variance in university GPA and *g* predicted a remarkable additional $\Delta R^2 = 0.402^2 = 16.0\%$ ($p = 0.003$).

In a second model (Model B, **Figure 2B**), we tested the incremental validity of high school GPA above and beyond

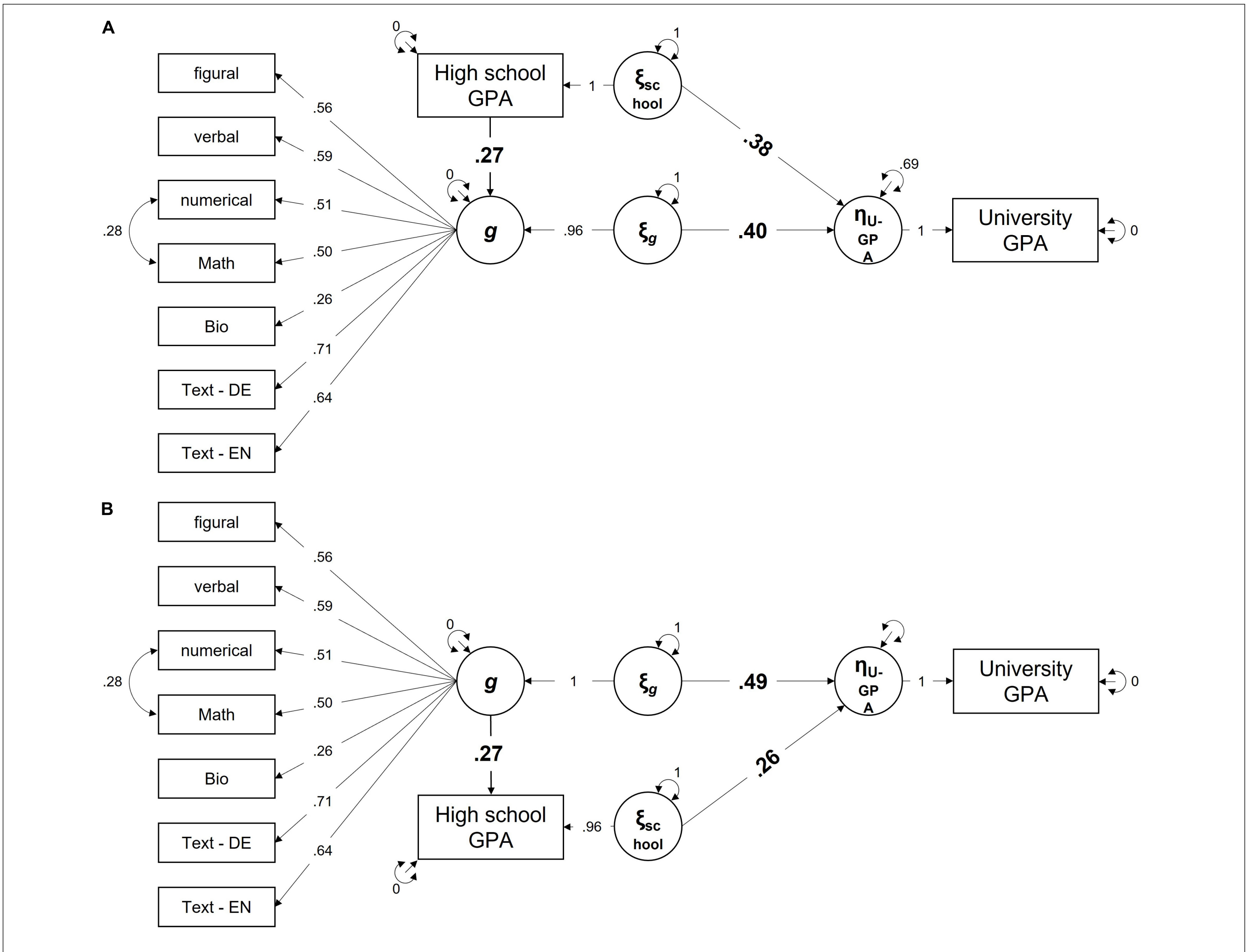


FIGURE 2 | Model-based test of incremental validity of **(A)** *g* over high school grade point average (GPA) and **(B)** high school GPA over *g*. $\chi^2_{(25)} = 40.68$, $p = 0.025$, CFI = 0.958, RMSEA = 0.041. All estimates are standardized and significantly greater than zero.

g. The approach was equivalent to model A except that high school GPA was regressed on *g* to obtain the residual variance of high school GPA after controlling for g. In this model, *g* predicted $R^2 = 0.492^2 = 23.8\%$ ($p < 0.001$) of the variance in university GPA and high school GPA predicted an additional $\Delta R^2 = 0.262^2 = 6.9\%$ ($p = 0.097$).

DISCUSSION

The goal of the present study was to develop and validate an admission test for the highly selective bachelor psychology program in Germany. We based the test on the extensive literature on subject-specific achievement tests, contemporary models of cognitive ability, and a comprehensive taxonomy of the bachelor psychology curriculum. The test battery meets psychometric requirements and, importantly, predicts university GPA 2.5 years later above and beyond high school GPA.

Evidence of Construct Validity

Unidimensional measurement models could be established for all tests, which provides evidence for their construct validity and legitimizes the aggregation of items (Little et al., 2002). At the higher level, a single general factor explained the observed data well. Contrary to theoretical expectations, a reliable factor of crystallized intelligence could not be established. As a power analysis revealed, this difference compared to Formazin et al. (2011) was likely due to the small and preselected sample of the current study. In future unselected applicant samples a reliable *gc* factor that incrementally contributes to the prediction of academic achievement might emerge. Similarly, the knowledge factor could also be strengthened by additional indicators, a point we address in the discussion on possible further developments of the test battery. Ultimately, however, this is an empirical question–for the time being, the single-factor model stands, which is in excellent agreement with the manifest total score of the test battery, which is to be used later for admission decisions.

Evidence of Criterion Validity

The test battery predicted university GPA 2.5 years later. Importantly, it was incrementally valid above high school GPA and vice-versa. Thus, both selection criteria significantly contributed to the prediction of academic achievement in the bachelor studies of psychology. In fact, the predictive power of the test was higher than that of high school GPA—while the test explained substantial portions of university GPA after controlling for high school GPA, high school GPA provided less incremental explanation after controlling for test performance. This speaks for the great utility of the test and the importance of cognitive abilities for university achievement.

Regarding high school GPA, it must be acknowledged that the range of grades was strongly restricted because participants were admitted to psychology studies based on their (very good) high school GPA before the study. This range restriction invariably attenuates observable associations between school and university GPA (Sackett and Yang, 2000). University GPA was also highly restricted in range, with most grades indicating that the ceiling for performance was too low. Whether this is due to uniformly exceptional psychology students or an inflationary allocation of good and very good grades can be discussed, but statistically it attenuates the observed criterion correlations for both the test and high school GPA. Criterion correlations might therefore be higher if standardized knowledge tests with adequate difficulty would be used to assess academic achievement (e.g., in the form of a master's admission test). In this case, lower estimates of predictive validity would be expected for high school GPA as it would not benefit from a mono method bias (Campbell and Fiske, 1959) anymore, as is now the case with GPA as the criterion.

Broadening the Predictor Space

The test represents a comprehensive measure of reasoning and to some extent knowledge and text comprehension which jointly predict academic achievement, but the inclusion of additional predictors might further improve its predictive validity. Given the meta-analytic evidence for the strong predictive validity of subject-specific knowledge tests (e.g., Kuncel et al., 2001) strengthening this component in the test battery might be envisioned in the future. In Austria, for example, extensive psychology knowledge tests, for which applicants can prepare in advance, are successfully applied (Legenfelder et al., 2008). Domain sampling [cf. Robitzsch (2015); chap. 7] based on the comprehensive taxonomy proposed in this study would allow for a content valid assessment of relevant prior knowledge that applicants would benefit from during their studies [e.g., Hambrick and Engle, 2002; but see also Simonsmeier et al. (2021), for a critical discussion on the effects of prior knowledge on learning].

Obviously, other attributes and abilities necessary to succeed in both psychology degrees and psychological occupations are not measured by the present test battery. These traits include conscientiousness, motivation, and other self-reported dispositions—the measurement of which is too easily distorted. In areas such as psychology, medicine, or teaching, socio-emotional abilities could be potent predictors of academic and job achievement and policymakers seem ready to include such abilities in college admission procedures. Yet, it is simple to demand such an extension of the predictor space, yet it is hard to scientifically justify. Most measures of socio-emotional abilities are severely plagued by psychometric issues, ranging from assessing maximal abilities with self-reports of typical behavior, over tests without veridical scoring, to measurement models not matching theoretical models (Wilhelm, 2005a). Although some approaches overcome these issues (Hildebrandt et al., 2015; Schlegel and Scherer, 2018; Geiger et al., 2021), a comprehensive model of socio-emotional abilities as a second stratum factor is still missing (Olderbak and Wilhelm, 2020). At the present moment, socio-emotional abilities are merely potential candidates to be included in comprehensive models of cognitive abilities (Wilhelm and Kyllonen, 2021) and are not yet suitable for use in university admissions. We hope that this gap will soon be filled to broaden the space of valid predictors.

Limitations and Future Perspectives

A limitation of the current study is the pre-selected nature of the sample. All participants were already admitted to psychology studies based on their (very good) high school GPA. Thus, the sample is likely not representative of prospective, unselected applicant samples. For the calibration of item difficulties, however, this was beneficial because the test will have to differentiate in the high ability spectrum.

Fairness is an important aspect of admission procedures. Albeit, it is also difficult to assess and address. The scientific concepts of fairness (American Educational Research Association [AERA] et al., 2014) predominantly deviate from common conceptions of group-equal representations because they allow for potentially real group differences in measured abilities. Instead, the American Educational Research Association [AERA] et al. (2014) stress (a) the equal treatment of all applicants, (b) the opportunity to familiarize oneself with the selection procedure for all applicants, and (c) the absence of systematic discrimination against groups. Standardized achievement tests undoubtedly meet (a and b) has been addressed for the current test through free provision of preparation materials. Testing for systematic discrimination of groups (e.g., age, gender, SES, ethnicity) is a high priority for the future. However, this presupposes sufficient data for statistical analysis. This was not the case in the preselected sample of our study, which was rather homogeneous in terms of gender (i.e., mostly female), age (i.e., young), and SES (i.e., mostly average to high). Thus, some aspects of fairness can only be investigated once the test is operational.

We argue that an objective, standardized, performance test as developed here has the potential to reduce unfairness and so far we have no reason to expect that the test violates scientific standards of faking. However, causes for group differences supposedly precede testing and what might appear as true group difference might in fact be due to unequal group treatment at an earlier stage. Regarding individual tests, one might fear that tests covering school knowledge are prone to effects of SES just like school grades are. School grades and knowledge tests differ in two aspects relevant for considering their fairness: standardized tests allow for a rather pure assessment of individual differences in declarative knowledge and their content is both tractable and transparent. In this sense, standardized knowledge tests supposedly allow for fairer comparisons between individuals.

Lastly, adherence to the paper-pencil testing in German university admission should be reconsidered. Computerizing the test would not only increase efficiency but also allow for the implementation of numerous alternative test formats, e.g., working memory capacity (WMC) (Wilhelm et al., 2013). Contrary to fluid intelligence or g, WMC has a much stronger theoretical foundation in cognitive psychology which greatly facilitates a number of psychometric challenges (e.g., easier or even automated item development, strong determination of item difficulties). Some tests from the present battery are available in a digital format and all could be digitized easily. Similar tests suggest equivalence of digital and paper-based versions (e.g., Schroeders and Wilhelm, 2010, 2012).

CONCLUSION

In sum, the psychology-specific admission test developed in this study meets psychometric requirements and predicts academic achievement beyond high school GPA. It thus contributes substantially to the identification of suitable psychology students and should be integrated into established admission procedures.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because of data privacy restrictions. Requests to access the datasets should be directed to LW, luc.watrin@uni-ulm.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethikkommission der Fakultät für Verhaltens- und Empirische Kulturwissenschaften Heidelberg. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LW: conceptualization, data curation, formal analysis, methodology, validation, visualization, and writing—original draft preparation. MG: conceptualization, methodology, validation, and writing—sections and revisions. JL: data curation, resources, review, and editing. BS: conceptualization, funding acquisition, project administration, review, and editing. OW: conceptualization, funding acquisition, project administration, supervision, and writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partly funded by the State of Baden-Württemberg, Ministry of Science, Research and the Arts (Fonds Erfolgreich Studieren in Baden-Württemberg—FESt-BW—2. Tranche Förderlinie 4 Eignung und Auswahl).

ACKNOWLEDGMENTS

We thank all project partners who were involved in the test administration and retrieval of criterion data at the different universities.

REFERENCES

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

American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (eds) (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Autorengruppe Bildungsberichterstattung (2020). Bildung in Deutschland 2020: Ein Indikatorengestützter Bericht Mit Einer Analyse zu Bildung in Einer Digitalisierten Welt. WBV Media. Available online at: https://www.bildungsbericht.de/de/bildungsberichte-seit-2006/bildungsbericht-2020/pdf-dateien-2020/bildungsbericht-2020-barrierefrei.pdf (accessed March 4, 2022).

Beard, J., and Marini, J. (2018). *Validity of the SAT® for Predicting First-Year Grades: 2013 SAT Validity Sample*. New York, NY: College Board.

Becker, N., Schmitz, F., Falk, A. M., Feldbrügge, J., Recktenwald, D. R., Wilhelm, O., et al. (2016). Preventing response elimination strategies improves the convergent validity of figural matrices. *J. Intelligence* 4:2. doi: 10.3390/jintelligence4010002

Bridgeman, B., McCamley-Jenkins, L., and Ervin, N. (2000). Predictions of freshman grade-point average from the revised and recentered SAT®I: reasoning test. *ETS Res. Rep. Ser.* 2000, i–16. doi: 10.1002/j.2333-8504.2000.tb01824.x

Bundesverfassungsgericht [BVerfG] (2017). *Urteil des Ersten Senats vom 19. Dezember 2017– 1 BvL 3/14 – Rn. (1 – 253)*. Available online at: http://www.bverfg.de/e/ls20171219_1bvl000314. html

Camara, W. J., and Kimmel, E. W. (eds) (2005). *Choosing Students: Higher Education Admissions Tools for the 21st Century*. Mahwah, NJ: L. Erlbaum Associates.

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56:81. doi: 10.1037/h0046016

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge: Cambridge University Press.

Cattell, R. B. (1987). *Intelligence: Its Structure, Growth and Action*. North Holland: Elsevier.

Coyle, T. R. (2006). Test–retest changes on scholastic aptitude tests are not related to g. *Intelligence* 34, 15–27. doi: 10.1016/j.intell.2005.04.001

Cronbach, L. J. (1960). *Essentials of Psychological Testing*, 2nd Edn. Hoboken, NJ: John Wiley & Sons.

Deutschen Gesellschaft für Psychologie [DGPs] (2022). Statut für die Vergabe des "Qualitätssiegels für Psychologische Bachelorstudiengänge an Deutschsprachigen Hochschulen" der Deutschen Gesellschaft für Psychologie (DGPs) [Statute for the Award of the "Seal of Quality for Psychological Bachelor's Programs at German-Language Universities" of the of the German Society for Psychology (DGPs)]. Available online at: https://zwpd.transmit.de/images/zwpd/dienstleistungen/qualitaetssiegel/statut_qualitaetssiegel_bachelor_2022-06-16.pdf (accessed June 23, 2022).

DIN (2016). *DIN 33430: Requirements for Proficiency Assessment Procedures and Their Implementation*. Berlin: Beuth.

Donovan, J. J., Dwight, S. A., and Schneider, D. (2014). The impact of applicant faking on selection measures, hiring decisions, and employee performance. *J. Bus. Psychol.* 29, 479–493. doi: 10.1007/s10869-013-9318-5

Estrada, E., Ferrer, E., Abad, F. J., Román, F. J., and Colom, R. (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: a longitudinal multi-group latent-variable study. *Intelligence* 50, 93–99. doi: 10.1016/j.intell.2015.02.004

Feng, Y., and Hancock, G. R. (2021). *Model-Based Incremental Validity. Psychological Methods*. Washington, DC: American Psychological Association.

Formazin, M., Schroeders, U., Köller, O., Wilhelm, O., and Westmeyer, H. (2011). Studierendenauswahl im Fach Psychologie: Testentwicklung und Validitätsbefunde. *Psychol. Rundschau* 62, 221–236. doi: 10.1026/0033-3042/a000093

Geiger, M., Bärwaldt, R., and Wilhelm, O. (2021). The good, the bad, and the clever: faking ability as a socio-emotional ability. *J. Intelligence* 9, 1–22. doi: 10.3390/jintelligence9010013

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annu. Rev. Psychol.* 60, 549–576. doi: 10.1146/annurev.psych.58.110405.085530

Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence* 8, 179–203. doi: 10.1016/0160-2896(84)90008-4

Hambrick, D. Z., and Engle, R. W. (2002). Effects of domain knowledge, working memory capacity, and age on cognitive performance: an investigation of the knowledge-is-power hypothesis. *Cogn. Psychol.* 44, 339–387. doi: 10.1006/cogp.2001.0769

Heene, M. (2007). *Konstruktion und Evaluation eines Studierendenauswahlverfahrens für Psychologie an der Universität Heidelberg*. Unpublished doctoral dissertation. Heidelberg: Universität Heidelberg.

Hell, B., Trapmann, S., Weigand, S., and Schuler, H. (2007b). Die Validität von Auswahlgesprächen im Rahmen der Hochschulzulassung—eine Metaanalyse. *Psychol. Rundschau* 58, 93–102. doi: 10.1026/0033-3042.58.2.93

Hell, B., Trapmann, S., and Schuler, H. (2007a). Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum. *Empirische Pädagogik* 21, 251–270.

Hildebrandt, A., Sommer, W., Schacht, A., and Wilhelm, O. (2015). Perceiving and remembering emotional facial expressions—A basic facet of emotional intelligence. *Intelligence* 50, 52–67. doi: 10.1016/j.intell.2015.02.003

International Organization for Standardization [ISO] (2011). *ISO 10667-2. Assessment Service Delivery—Procedures and Methods to Assess People in Work and Organizational Settings—Part 2: Requirements for Service Providers*. Geneva: ISO.

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen. Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells [Multimodal classification of intelligence performance. Experimentally controlled development of a descriptive intelligence structure model]. *Diagnostica* 28, 195–226.

Kersting, M. (2008). Zur Akzeptanz von Intelligenz- und Leistungstests. *Rep. Psychol.* 33, 420–433.

Koenig, K. A., Frey, M. C., and Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence* 36, 153–160. doi: 10.1016/j.intell.2007.03.005

Köller, O., Watermann, R., Trautwein, U., and Lüdtke, O. (2004). "Wege zur Hochschulreife in Baden-Württemberg — Erweiterung von Bildungswegen und Studiereignung: die grundlegenden Fragestellungen in TOSCA," in *Wege zur Hochschulreife in Baden-Württemberg*, eds O. Köller, R. Watermann, U. Trautwein, and O. Lüdtke (Wiesbaden: VS Verlag für Sozialwissenschaften), 113–119. doi: 10.1007/978-3-322-80906-3_5

Kulik, J. A., Bangert-Drowns, R. L., and Kulik, C.-L. C. (1984). Effectiveness of coaching for aptitude tests. *Psychol. Bull.* 95, 179–188. doi: 10.1037/0033-2909.95.2.179

Kuncel, N. R., Hezlett, S. A., and Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: implications for graduate student selection and performance. *Psychol. Bull.* 127, 162–181. doi: 10.1037/0033-2909.127.1.162

Kunina, O., Wilhelm, O., Formazin, M., Jonkmann, K., and Schroeders, U. (2007). Extended criteria and predictors in college admission: exploring the structure of study success and investigating the validity of domain knowledge. *Psychol. Sci.* 49, 88–114.

Legenfelder, P., Baumann, U., Allesch, C., and Nürk, H. C. (2008). "Studierendenauswahl an der Universität Salzburg: Konzeption und Validität," in *Studierendenauswahl und Studienentscheidung*, eds H. Schuler and B. Hell (Göttingen: Hogrefe).

Levacher, J., Koch, M., Hissbach, J., Spinath, F. M., and Becker, N. (2022). You can play the game without knowing the rules – but you're better off knowing them: the influence of rule knowledge on figural matrices tests. *Eur. J. Psychol. Assess.* 38, 15–23. doi: 10.1027/1015-5759/a000637

Little, T. D., Cunningham, W. A., Shahar, G., and Widaman, K. F. (2002). To parcel or not to parcel: exploring the question, weighing the merits. *Struct. Equ. Model.* 9, 151–173. doi: 10.1207/S15328007SEM0902_1

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., and Maurer, S. D. (1994). The validity of employment interviews: a comprehensive review and meta-analysis. *J. Appl. Psychol.* 79, 599–616. doi: 10.1037/0021-9010.79.4.599

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Erlbaum.

Olderbak, S., and Wilhelm, O. (2020). Overarching principles for the organization of socioemotional constructs. *Curr. Direct. Psychol. Sci.* 29, 63–70. doi: 10.1177/0963721419884317

Perfetto, G., Escandón, M., Graff, S., Rigol, G., and Schmidt, A. (1999). *Toward a Taxonomy of the Admissions Decision-Making Process: A Public Document Based on the First and Second College Board Conferences on Admissions Models*. New York, NY: College Entrance Examination Board.

Postlethwaite, B. E. (2011). *Fluid Ability, Crystallized Ability, and Performance Across Multiple Domains: A Meta-Analysis*. Ph.D. thesis. Iowa City, IA: University of Iowa.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Revelle, W. (2020). *Psych: Procedures for Psychological, Psychometric, and Personality Research (R package Version 2.0.12)*. Available online at: https://cran.r-project.org/web/packages/psych/psych.pdf (accessed February 22, 2021).

Robitzsch, A. (2015). *Essays zu Methodischen Herausforderungen im Large-Scale Assessment*. Unpublished doctoral dissertation. Berlin: Humboldt-Universität zu Berlin.

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Sackett, P. R., and Yang, H. (2000). Correction for range restriction: an expanded typology. *J. Appl. Psychol.* 85, 112–118. doi: 10.1037/0021-9010.85.1.112

Schipolowski, S., Wilhelm, O., and Schroeders, U. (2014). On the nature of crystallized intelligence: the relationship between verbal ability and factual knowledge. *Intelligence* 46, 156–168. doi: 10.1016/j.intell.2014.05.014

Schlegel, K., and Scherer, K. R. (2018). The nomological network of emotion knowledge and emotion understanding in adults: evidence from two new performance-based tests. *Cogn. Emot.* 32, 1514–1530. doi: 10.1080/02699931.2017.1414687

Schmidt-Atzert, L. (2005). Prädiktion von Studienerfolg bei Psychologiestudenten. *Psychol. Rundschau* 56, 131–133. doi: 10.1026/0033-3042.56.2.131

Schneider, W. J., and McGrew, K. S. (2018). "The Cattell–Horn–Carroll model of cognitive abilities," in *Contemporary Intellectual Assessment*, 4th Edn, eds D. P. Flanagan and E. M. McDonough (New York, NY: The Guilford Press), 73–163.

Schroeders, U., and Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *Eur. J. Psychol. Assess.* 26, 284–292. doi: 10.1027/1015-5759/a000038

Schroeders, U., and Wilhelm, O. (2012). Equivalence of reading and listening comprehension across test media. *Educ. Psychol. Meas.* 71, 849–869. doi: 10.1177/0013164410391468

Schult, J., Hofmann, A., and Stegt, S. J. (2019). Leisten fachspezifische Studierfähigkeitstests im deutschsprachigen Raum eine valide Studienerfolgsprognose?: ein metaanalytisches Update. *Z. Entwicklungspsychol. Pädagog. Psychol.* 51, 16–30. doi: 10.1026/0049-8637/a000204

Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., and Schneider, M. (2021). Domain-specific prior knowledge and learning: a meta-analysis. *Educ. Psychol.* 57, 31–54. doi: 10.1080/00461520.2021.1939700

Tippins, N. T., Sackett, P., and Oswald, F. L. (2018). Principles for the validation and use of personnel selection procedures. *Industrial Organ. Psychol.* 11, 1–97. doi: 10.1017/iop.2018.195

Trapmann, S., Hell, B., Weigand, S., and Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs—eine Metaanalyse. *Z. Pädagog. Psychol.* 21, 11–27. doi: 10.1024/1010-0652.21.1.11

Viswesvaran, C., and Ones, D. S. (1999). Meta-analyses of fakability estimates: implications for personality measurement. *Educ. Psychol. Meas.* 59, 197–210. doi: 10.1177/00131649921969802

Voyer, D., and Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. *Psychol. Bull.* 140, 1174–1204. doi: 10.1037/a0036620

Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M. R., and Schmidt, F. L. (2015). College performance and retention: a meta-analysis of the predictive validities of ACT® scores, high school grades, and SES. *Educ. Assess.* 20, 23–45. doi: 10.1080/10627197.2015.997614

Wetzenstein, E. (2004). Entwicklung eines Anforderungsprofils für Studierende am Beispiel der Psychologie [Development of Job Specifications for Students in Psychology]. Paper presented at the Bühler-Kolloquium of the TU-Dresden. Dresden: Technische Universität Dresden.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019). Welcome to the Tidyverse. *J. Open Source Softw.* 4:1686. doi: 10.21105/joss.01686

Wilhelm, O. (2005b). "Measuring reasoning ability," in *Handbook of Understanding and Measuring Intelligence*, eds O. Wilhelm and R. W. Engle (Thousand Oaks, CA: SAGE Publications, Inc), 373–392. doi: 10.4135/9781452233529.n21

Wilhelm, O. (2005a). "Measures of emotional intelligence: practice and standards," in *Emotional Intelligence: An International Handbook*, eds R. Schulze and R. D. Roberts (Göttingen: Hogrefe), 131–154.

Wilhelm, O., Hildebrandt, A., and Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Front. Psychol.* 4:433. doi: 10.3389/fpsyg.2013.00433

Wilhelm, O., and 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. doi: 10.1016/j.intell.2021.101585

Yu, C.-Y. (2002). *Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes*. Doctoral dissertation. California, CA: University of California.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer MK declared a shared affiliation with the author JL to the handling editor at the time of review.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX

Appendix Table 1 | Simple and multiple linear regression for predicting university grade point average (GPA) with (a) high school GPA, (b) test score or both.

Predictors	**Beta [CI$_{95\%}$]**	R^2	ΔR^2
(a) High school GPA			
Step 1		12.4%	
(Intercept)	0.00 [–0.12, 0.13]		
High school GPA	0.37 [0.24, 0.51]		
Step 2		21.1%	8.7%
(Intercept)	0.04 [–0.08, 0.16]		
High school GPA	0.31 [0.18, 0.44]		
Test	0.30 [0.18, 0.43]		
(b) Test score or both			
Step 1		12.9%	
(Intercept)	0.04 [–0.09, 0.17]		
Test	0.36 [0.23, 0.49]		
Step 2		21.1%	8.2%
(Intercept)	0.04 [–0.08, 0.16]		
Test	0.31 [0.18, 0.44]		
High school GPA	0.30 [0.18, 0.43]		

Beta, standardized regression coefficient; CI$_{95\%}$, 95% confidence interval.