Polskie Forum Psychologiczne, 2023, tom 28, numer 1, s. 79–98
DOI: 10.34767/PFP.2023.01.05

DEVELOPMENT AND VALIDATION OF THE SMAT-G COGNITIVE ABILITIES TEST*

Jarosław Grobelny[1]

Summary. The presented paper describes the development of a new tool for measuring cognitive abilities: SMAT-G. The test is based on the Cattell–Horn–Carroll model of intelligence (i.e., the three-stratum structure of abilities). The scale selection (including fluid and crystallized abilities, reading and comprehension, and quantitative knowledge) and development of test items are presented. Three validation studies were conducted. These studies analyzed response processes, item difficulty and discrimination, split-half reliability, factor structure, and test consequences. SMAT-G results fit the theoretical model and presented significant correlations with recognized tests of cognitive abilities. The results confirm the reliability and validity of SMAT-G.

Key words: cognitive abilities test, specific mental abilities, general mental ability, fluid intelligence, crystallized intelligence, testing

Introduction

Cognitive abilities testing has always been a major driving force for both practice and research in the field of psychology. For example, in industrial and organizational psychology, studies utilizing the measurement of cognitive ability have contributed to the understanding of job performance, one of the most important criteria investigated in the field (Rojon, McDowall, Saunders, 2015). As a result, cognitive ability is considered one of the critical individual factors driving job performance and career success (Lang, Kell, 2019; Sackett et al., 2021). Cognitive abilities are thought to contribute to performance by either supporting employees' acquisition of tacit knowledge or by working as a meta-component responsible

* Disclosure statement: The authors report there are no competing interests to declare.

[1] Wydział Psychologii i Kognitywistyki, Uniwersytet im. Adama Mickiewicza w Poznaniu (Faculty of Psychology and Cognitive Science, Adam Mickiewicz University in Poznan), ORCID: 0000-0003-4296-402X.

Mailing address: Jarosław Grobelny,
jaroslaw.grobelny@amu.edu.pl

for identifying the encountered work task-related issue and choosing a strategy to solve it (Wagner, Sternberg, 1987; Schmidt, Hunter, 1993; Grobelny, 2018). Cognitive abilities support complex work-related problem-solving in particular, as suggested by the results of meta-analyses demonstrating substantial predictive validity of this attribute in high-complexity occupations in various cultural and occupational contexts (Salgado et al., 2003; Salgado, 2017).

Even though cognitive testing has over a century of history and numerous tests are available, the introduction of a new tool can still be justified theoretically and practically. There is an ongoing and heated debate over whether general or specific cognitive abilities are superior predictors of job performance (see Kell, Lang, 2018). Some researchers have expressed serious concerns about whether general mental ability (GMA), the core construct measured by intelligence tests, is actually a leading predictor of employee behavior and performance (Richardson, Norgate, 2015; Grobelny, 2018). Practitioners need tools to enhance personnel decision-making in a labor market with a workforce shortage. In addition, it was determined that there is a shortage of tools that meet the following criteria: are based on a robust theory of intelligence and present psychometric properties that meet appropriate standards; allow estimated results for general ability and more detailed factors; are published under an open or permissive license that allows users to adapt it to their own needs; allow group and remote testing; and, last but not least, are suitable for application in a work context. Aforementioned reasons, among several others, justify further work to improve ability testing. The presented paper describes the initial development of the Specific Mental Abilities Test for general and occupational samples (SMAT-G) content and the results of a series of validity studies that proved it to be a reliable and valid test of cognitive abilities.

Theoretical Framework

The Cattell–Horn–Carroll model (hereinafter referred to as CHC) was chosen as the theoretical foundation for the developed test, as it is considered a leading psychometric theory of human abilities (Alfonso, Flanagan, Radwan, 2005; Kaufman, 2009; Schneider, Newman, 2015). CHC states that human intelligence has a complex hierarchical structure with three levels called *strata* (Schneider, McGrew, 2012). At the top level (stratum III), there is a single general factor that is considered to be a unitary construct; according to Carroll (1993), this factor is assumed to account for the correlations between broad abilities. Next, broad abilities (stratum II) are defined as core and long-standing characteristics of individuals that impact their behaviors in a particular domain to a substantial degree (Carroll, 1993, p. 634). With each version of the CHC model, the number of broad abilities increased (from the original eight up to the 17 listed currently); however, only ten of them are widely accepted (Flanagan, Ortiz, Alfonso, 2007). This list includes fluid reasoning (Gf), short-term/primary memory (Gsm), long-term storage and retrieval (Glr), processing speed

(Gs), reaction and decision speed (Gt), psychomotor speed (Gps), comprehension and knowledge (Gc), domain-specific knowledge (Gkn), reading and writing (Grw), quantitative knowledge (Gq), visual processing (Gv) and auditory processing (Ga). Schneider himself admitted that the descriptions of the broad abilities incorporated in later revisions of CHC are not robust enough due to the lack of independent replications (Schneider, Newman, 2015). Each broad ability includes a large number of narrow abilities (stratum III) that "represent greater specializations of abilities, often in quite specific ways that reflect the effects of experience and learning, or the adoption of particular strategies of performance" (Carroll, 1993, p. 634). More than 70 of them have been described to date (Schneider, Newman, 2015), including inductive, deductive, and quantitative reasoning, memory span, association fluency, simple reaction time, general verbal information, lexical knowledge, reading decoding, mathematic knowledge, and many others.

SMAT-G Development

The first step in SMAT-G development was to choose the abilities to be measured by the test. This was done with the following underlying assumptions: the scope of the test should include at least four broad abilities, with both groups from Cattell and Carroll's original model; it should allow self-administered, remote, and group testing; the chosen abilities should be testable without sophisticated technological support (e.g., chronometry measurement or physiological tests) and in a single testing session. After analysis of CHC's stratum II abilities, the following were designated: Gf, Gc, Grw, and Gq. Gf and Gc are both essential parts of intelligence structure and are high on *g* loading, thus are a must-have in any cognitive test (Schneider, McGrew, 2018). Grw was chosen due to the potential impact on behavior and outcomes in the workplace, as written communication and presentations are often included in job performance models (Viswesvaran, Ones, 2000). Likewise, due to the high demand for numeric data analysis to enhance professional decision-making by employees nowadays, Gq was chosen.

Next, the narrow abilities which would form the direct basis of the SMAT-G scales were chosen. The critical requirement during this process was to include narrow abilities, which, according to Schneider and McGrew, are a core part of their broad groups, such as inductive reasoning for Gf (Schneider, McGrew, 2012). Psychometric tools, test kits, and works provided by recognized authors (with open and permissive licenses) were used as the sources of the specific cognitive task types (Stankov, 1997; Learning Express, 2005; Sternberg, Kaufman, Grigorenko, 2008; Nisbett, 2009; Sourceforge, 2012; Condon, Revelle, 2014; Cambridge Brain Sciences, 2017; Open Source Psychometrics Project, 2017a, 2017b). Table 1 presents the results of this early development, i.e., a selection of test scales with their names and corresponding abilities from CHC and a basic description of the CHC model components based on Schneider and McGrew (2012).

Table 1. Selected broad and narrow abilities described in the CHC model and SMAT-G initial version scales based on these abilities

CHC model selected components		SMAT-G scales and their corresponding narrow abilities		
Broad ability name	Broad ability definition	Scale	Scale name	Narrow ability tested by the scale
Fluid reasoning (Gf)	The control of attention (flexible and deliberate) to solve novel problems that cannot be addressed by past habits, schemas, or prior learning. Includes narrow abilities such as induction (I), deductive reasoning (RG), quantitative reasoning (RQ).	Gf_1	Relations between pairs	Induction (I)
		Gf_2	Series of numbers	Quantitative reasoning (RQ)
		Gf_3	Syllogisms	Deductive reasoning (RG)
		Gf_4	Analogies	Induction (I)
Reading and writing (Grw)	The depth and scope of knowledge and skills related to written language and reading. Includes narrow abilities such as reading decoding (RD), reading comprehension (RC), spelling ability (SG).	Grw_1	Matching definitions	Reading comprehension (RC)
		Grw_2	Questions to a story	Reading comprehension (RC)
		Grw_3	Summaries	Reading comprehension (RC)
		Grw_4	Questions to a story (complex)	Reading comprehension (RC)
Comprehen-sion-Know-ledge (Gc)	The depth and scope of knowledge and skills (valued by one's culture) acquired by an individual. Includes narrow abilities such as general verbal information (K0), lexical knowledge (VL).	Gc_1	Terms grouping	General verbal information (K0)
		Gc_2	Definitions of concepts	General verbal information (K0)
		Gc_3	Synonyms and antonyms	Lexical knowledge (VL)
Quantitative knowledge (Gq)	The depth and scope of knowledge explicitly related to mathematics (but not proficiency in performing calculations). Includes narrow abilities such as mathematical knowledge (KM) and achievements (A3).	Gq_1	Test of mathematical knowledge	Mathematical knowledge (KM)

The development and validation process was based on both APA and EFPA standards for psychological testing (American Educational Research Association et al., 1999; European Federation of Psychologists' Associations, 2013). According to APA, one source of validity evidence is based on response processes; therefore, selected cognitive tasks and their underlying processes were analyzed in the initial step. A positive outcome of this a priori analysis was intended to provide greater confidence that the characteristics tested by the scales would correspond to the relevant narrow abilities.

Responses Process Analysis

In Gf_1, the task is to pair a keyword with one out of four presented words so that they would be linked by the same relation that a provided key pair. One must discover this relation oneself; therefore, induction reasoning must be applied. In Gf_2, the task is to analyze a series of numbers and then fill in one or two spaces to complete the series following the underlying rule. Participants must figure out these rules themselves by using inductive reasoning and conducting a series of computations (incl. addition, subtraction, and multiplication, applied to single or multiple sequences simultaneously); this complies with the definition of quantitative reasoning. When presenting the figures, no context was provided to focus the participants' attention on the numerical material. Syllogism (Gf_3) is a popular exercise in which one is provided with a series of assumptions. Then, based only on these assumptions, one must decide whether a statement is correct or not (or if this cannot be judged). As reasoning is conducted based on a known premise, deductive reasoning is required from the person solving this test. In Gf_4, one must solve an analogy in the following form: A is to B as C is to D. However, four possibilities are provided for D, and inductive reasoning must be applied. In Grw_1, a definition of a phenomenon of a legislative or economic nature (to reduce the involvement of acquired knowledge) is provided, followed by four descriptions of short situations. The task is to indicate which descriptions fit the provided definition using only the received text. This requires the ability to focus on appropriate words and their meanings; therefore, this task matches the definition of reading comprehension. Grw_2 presents a short story supplemented with four statements about it. Only one is correct (its correctness can be determined from the text itself), and the participant must choose it. Once again, success in this task relies on formal understating of certain words and statements, so comprehension is necessary. In Grw_3, a short story is described, and participants are asked to choose from five available responses, one of which best reflects the story's theme. As this task requires analysis and understanding of the provided text, reading comprehension ability must be applied. Grw_4 is comparable to Grw_2, but this time a slightly longer text is provided with four supplementary statements, each of which could be true or false (or undetermined). Reading comprehension must be employed as one must rely on formal analysis and understanding of exact words. In

Gc_1, a keyword (real-life objects or phenomena) and four pairs of terms are provided. Based on one's knowledge, the task is to choose a pair to link all three words. Unlike Gf_1 or Gf_4, participants are asked to refer to their understanding of the phenomenon. Providing a correct answer requires general verbal information. The same ability is needed in Gc_2, in which a keyword is provided along with four possible words. One must choose the one that is an indispensable and definite part of a given keyword. Once again, respondents must refer to their acquired knowledge. In Gc_3, a series of words is provided, and one must choose synonyms for the first half of the list and antonyms for the second half. Success in this task depends on knowledge of definitions of words, which requires lexical knowledge. Finally, Gq_1 is a single-choice knowledge test from the mathematics domain, which directly represents mathematical knowledge (participants were asked, e.g., the definition of mathematical concepts or terms, characteristics of structures or geometric figures, and others; in each question, one correct and three incorrect responses were given). This a priori analysis indicated that the successful solving of the prepared tests may indeed require the assumed narrow abilities described by the CHC model, thus providing (certainly limited) evidence for its validity based on response processes.

Methods: Study 1 – Examination of the Initial Test Version

The first study aimed to evaluate the scales and items included in the initial version of SMAT-G and choose candidates for the final version.

Sample Description

There were 156 participants, 105 women (67.3%) and 51 men (32.7%); their average age was 24.62 (*SD* = 6.03). Of these participants, 83 were employed (53.2%) and had on average 4.42 years of professional experience (*SD* = 6.36) in various occupations. The respondents had the following education: vocational (7 people, 4.7%), high school (82 people, 55%), bachelor's degree (18 people, 12.2%), master's, and higher (42 people, 28.2%).

Measurements

The initial version of SMAT-G was the only psychometric tool used in the first study. It consisted of 12 scales, with 157 items in total. The test itself was delivered with the following instructions for participants: all questions are single-choice; avoid guessing as incorrect answers will be scored negatively; there is no time limit; the test should be completed with no technological support in a quiet environment and, ideally, with no breaks; if breaks are needed, they should be taken between scales, not in the middle of them. Additionally, the introduction encouraged participants to write their comments.

Procedure

Participants were recruited by the research support team using an opportunity sampling approach. After being informed about the study's purpose and anonymity, subjects were asked to sign a participation agreement. They were told that there would be no feedback from their results as the test was in an early development stage; this was also done to limit their motivation to guess or cheat. Each participant was handed an envelope with a printed copy of SMAT-G. They completed the test in their homes or places of their choosing. Then they were asked to return the envelope to the research team member within a week. Out of 194 envelopes returned, 156 contained completed tests.

Results: Study 1

Table 2 presents a proportion of correct answers for every item (which effectively was the item-difficulty index). Because none of the scales were in their final form, no analysis was possible on a scale level (e.g., item discrimination analysis). A qualitative review of participants' comments was carried out. According to APA guidelines, such feedback might serve as evidence of validity based on test content; thus, this part of the process needed to be carried out. Both quantitative and qualitative analyses were used to exclude inadequate items from the test. Items with too low or high scores (below 22% or above 90%) were removed. Items with three or more negative comments were also removed. Also, a few scales were completely excluded: Gf_4, Grw_1, Grw_2, and Gc_1. In Gf_4, two items did not meet the cut-off criteria, and respondents complained about three further items. As Gf_1 was also designed to measure induction, a decision was made to remove Gf_4. Because two items were removed from Grw_1, this scale would have been too short to differentiate respondents' characteristics. The same applied to Grw_2. Gc_1 was excluded due to numerous participants' comments. Finally, to limit the test length, the decision was made to exclude the entire part of Gc_4 with antonyms, as multiple questions were excluded from this scale. Overall, eight scales and 72 items were selected for further work. It was only after these questions were excluded that subsequent studies were conducted.

Methods: Study 2 – Examination of the Structure of Test Results

During the next study, SMAT-G was administered on a larger sample to ensure the robustness of analyses. Because items were dichotomous and there was only a single measurement, the split-half method was employed to assess test reliability (with odd and even items compared); therefore, it was expected that:

H1. The two halves of every scale will positively correlate with one another.

Confirmatory structure analysis was performed to verify whether the test results aligned with the structure of human intelligence described by CHC.

The aim was to provide evidence based on internal structure (following APA guidelines) and construct validity (by EFPA); therefore, it was expected that:

H2. The test results will fit the three-stratum structure described by CHC.

Information about another source of validity was gathered, namely testing consequences (hiring decisions in this context). Therefore, the results were compared between men and women. The between-group differences should be taken as a major indication that the test needs to be revised and rewritten, as its use in the recruitment context could lead to systematic inequality in hiring decisions.

Sample Description

Seven hundred ninety-seven people participated in the second study, of whom 500 (62.7%) were women and 291 (36.5%) were men. Six people did not provide this information. On average, participants were 31.85 years old (*SD* = 10.33, from 18 to 73). The majority, 533 people (66.9%), were employed, with 12.02 years of professional experience (*SD* = 9.98, from 3 months to 46 years). The sample was diverse in terms of education: 305 people had completed middle school (38.3%), 17 had completed technical or vocational school (2.1%), 185 had a bachelor's degree (23.2%), 50 had an engineer's degree (6.3%), 216 had a master's degree (27.1%), 21 had a master's in engineering (2.6%), and three had a Ph.D. (0.4%). Participants represented a broad scope of education fields.

Measurements

Only demographic and SMAT-G (limited to eight scales and 72 items) were measured. The overall result, namely General Mental Ability (GMA), was calculated as the mean for all eight scales.

Procedure

Participants were recruited by two methods: first, with the help of a support team that searched for subjects; second, with the help of twenty-one companies that forwarded the study invitation to their employees. Subjects were informed about the research purpose and the procedure. If they agreed to participate, they were provided with the individual access code and the link to the online form with SMAT-G. The test was supplemented by the same instruction and demographic questions as in Study 1. Subjects were asked to complete the test on their own, without using supportive devices, and preferably in a single session. Participants were told to take the test within a week, in the place and time of their choosing. Out of 960 invitations sent, 797 tests were completed; only 18 people started the procedure but did not finish it.

Table 2. Descriptive statistics for SMAT-G items from Study 1 and Study 2

Scale	Item	%*	%**	*Φ*
Gf$_1$	1	66%	-	-
	2′	68%	71%	.65
	3′	49%	37%	.68
	4′	23%	11%	.61
	5′	71%	69%	.60
	6′	42%	48%	.58
	7	68%	-	-
	8′	81%	49%	.56
	9′	56%	56%	.67
	10′	60%	53%	.60
	11	13%	-	-
Gf$_2$	1	95%	-	-
	2	88%	-	-
	3	84%	-	-
	4	94%	-	-
	5′	67%	65%	.75
	6′	81%	73%	.73
	7	93%	-	-
	8	76%	-	-
	9′	43%	26%	.66
cont. Gf$_2$	10′	46%	49%	.76
	11′	50%	49%	.82
	12′	71%	65%	.71
	13′	76%	64%	.72
	14′	13%	15%	.67
	15′	65%	53%	.73
	16′	16%	20%	.58
	17	3%	-	-
	18	10%	-	-
Gf$_3$	1	83%	-	-
	2	4%	-	-
	3	4%	-	-
	4′	84%	73%	.71
	5′	85%	85%	.61
	6	26%	-	-
	7′	60%	43%	.64
	8	22%	-	-
	9′	90%	80%	.76
	10′	79%	78%	.64
	11′	81%	83%	.76
cont. Gf$_3$	12	96%	-	-
	13	22%	-	-
	14	16%	-	-
	15′	46%	59%	.64
Grw$_3$	1	10%	-	-
	2′	83%	78%	.75
	3′	49%	24%	.67
	4′	62%	47%	.58
	5′	62%	51%	.65
	6′	51%	44%	.81
	7	78%	-	-
Grw$_4$	1′	70%	75%	.64
	2′	71%	68%	.71
	3′	70%	61%	.68
	4′	74%	62%	.66
	5	65%	-	-
	6	26%	-	-
	7	15%	-	-
	8	44%	-	-
	9′	30%	34%	.78

cont. Table 2

Scale	Item	%*	%**	Φ	Scale	Item	%*	%**	Φ	Scale	Item	%*	%**	Φ
cont. Grw$_4$	10′	67%	55%	.61	Gc$_3$	1′	34%	45%	.56	cont. Gq$_1$	7′	34%	43%	.62
	11′	81%	70%	.49		2′	74%	69%	.68		8′	58%	41%	.55
	12′	90%	86%	.64		3′	78%	72%	.64		9	6%	-	-
	13′	60%	59%	.63		4′	72%	81%	.82		10	14%	-	-
	14′	69%	68%	.69		5′	53%	59%	.69		11′	87%	74%	.58
	15′	38%	38%	.58		6′	40%	30%	.44		12′	49%	45%	.71
	16′	48%	34%	.49		7′	74%	77%	.70		13′	55%	53%	.78
Gc$_2$	1′	40%	42%	.57		8′	82%	76%	.67		14	36%	-	-
	2′	28%	15%	.64		9′	90%	86%	.70		15	24%	-	-
	3′	37%	38%	.72		10′	90%	87%	.59		16	75%	-	-
	4	29%	-	-		11′	43%	45%	.55		17	19%	-	-
	5′	53%	55%	.67		12′	69%	63%	.55		18′	80%	76%	.66
	6	43%	-	-	Gq$_1$	1	61%	-	-		19	24%	-	-
	7′	76%	78%	.49		2	40%	-	-		20	73%	-	-
	8′	39%	34%	.58		3	91%	-	-		21′	42%	35%	.80
	9′	61%	47%	.65		4′	58%	66%	.72					
	10′	38%	35%	.77		5′	15%	22%	.51					
	11	63%	-	-		6′	75%	69%	.69					

Scales Gf$_4$, Grw$_1$, Grw$_2$, Gc$_1$, and Gc$_4$ (antonyms) were removed from the test entirely and thus are excluded from the table. *Annotation.* ′ – item designated for further development; %* – the proportion of correct answers in Study 1; %** – the proportion of correct answers in Study 2; Φ – correlation coefficients between an item and a scale (with considered item excluded) in Study 2.

Results: Study 2

Table 2 presents descriptive statistics for SMAT-G items. Besides the means (which served as item-difficulty indexes) and deviations, item discrimination power indexes were computed using a correlation approach and *Φ* coefficients. On average, items were moderately correlated with their scales (*M* = .65, *SD* = .08); no coefficient was lower than .44, and the highest one was .82. All items met the goodness criteria based on Ebel & Frisbie's (1986) recommendations (*Φ* ≥ .40). Reported results proved the test construct's validity according to EFPA standards. Table 3 summarizes analyses on a scale level, with split-half reliability coefficients (with Spearman-Brown correction) and Shapiro-Wilk's test results included. Overall, these data indicate that items were correctly selected and that the appropriate differentiation of participants based on their results is possible. The item-difficulty indexes were relatively consistent. Most of them were in the 40–60% range; notably, most scales included both easier and slightly more difficult items. The strong relationship of the items to their scales and the high half-split reliability coefficients (above .80 in the majority of scales) proved the reliability of the entire test, thus confirming H1. All this consequently allowed further analysis of the tool's validity through CFA.

Before CFA was carried out, a correlation matrix of all test results was calculated. This matrix, presented in Table 3, indicates that a positive manifold, a fundamental assumption for abilities testing, had been met. This result alone provided evidence for test validity based on test content, as stated in APA guidelines. All the SMAT-G tests correlated positively and substantially; this is of utmost importance as the described phenomenon is often referred to as *the first law of intelligence* (Guttman, Levy, 1991). Thus, using R's *lavaan* package (Rosseel, Jorgensen, 2020), CFA was conducted to assess compliance between the actual data structure and the theoretical CHC model.

First, a test was performed to check the suitability of the data for structure detection. The KMO measure of sampling adequacy was as high as .84, and Bartlett's test of sphericity result indicated that CFA might be useful (χ^2 = 1135.91; *df* = 28; *p* < .001). Therefore, the following model was tested: results of narrow ability tests (I stratum) are loaded by factors corresponding to their broad abilities (II stratum), which in turn are loaded by a general factor (III stratum). Based on the theoretical premises, none of these factors were considered orthogonal. The standardized variance method was employed to estimate loading values and diagonally weighted least squares and calculate the parameter estimates. No missing data imputation method was applied because all missing data were counted as wrong answers. Figure 1 presents the data structure with latent factor loadings for broad and narrow abilities measures and error variance.

Table 3. Scale-level analysis of SMAT-G and *r* Pearson's correlations between scales

Scale	*M*	*SD*	*W*	*t*	Λ	Gf_1	Gf_2	Gf_3	Grw_3	Grw_4	Gc_2	Gc_3	Gq_1
Gf_1	49%	22%	.71	–.17	.74**	-							
Gf_2	48%	27%	.69	.77	.89***	.37***	-						
Gf_3	71%	24%	.78	.39	.80***	.38***	.40***	-					
Grw_3	49%	25%	.65	–1.02	.81***	.31***	.27***	.33***	-				
Grw_4	61%	21%	.63	–.47	.82***	.38***	.32***	.45***	.34***	-			
Gc_2	43%	23%	.71	–.98	.82***	.21***	.15***	.23***	.14***	.17***	-		
Gc_3	66%	21%	.58	1.17	.83***	.29***	.29***	.27***	.29***	.36***	.17***	-	
Gq_1	52%	24%	.63	.69	.79***	.24***	.36***	.24***	.11**	.25***	.14**	.34***	-
GMA	54%	14%	.76	.08	.95***	.65***	.63***	.69***	.58***	.65***	.53***	.58***	.51***

Annotation. *M* – average proportion of correct answers; *SD* – standard deviation for the average proportion of correct answers; *W* – results of Shapiro-Wilk of SMAT-G scales normality of distribution; *t* – *t*-test result for differences between man and women SMAT-G scales results; λ – corrected split-half reliability coefficient; * $p < .05$; ** $p < .01$; *** $p < .001$.

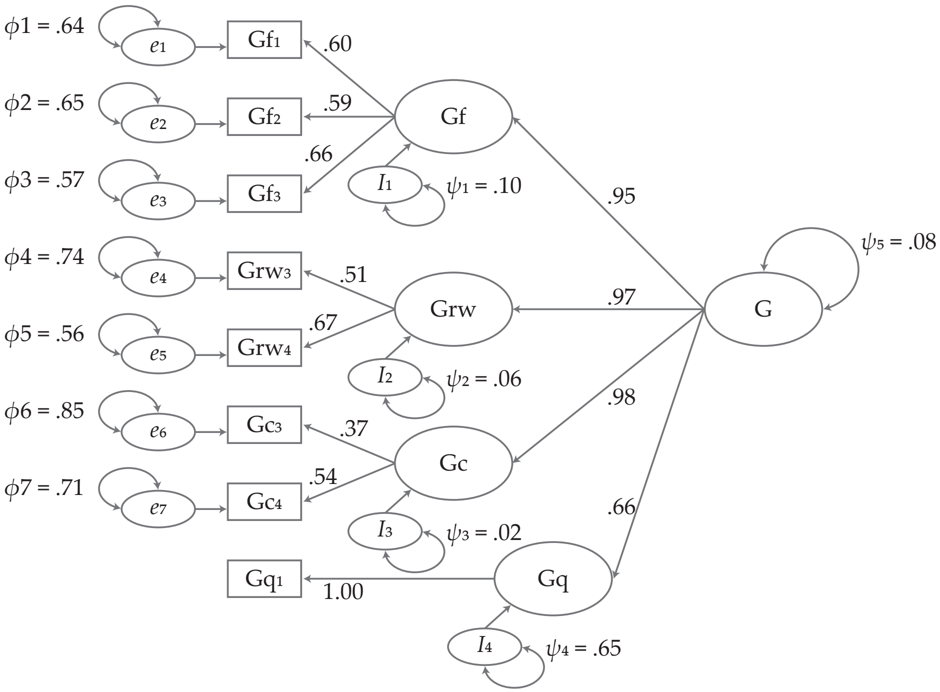

Figure 1. SMAT-G results structure with loadings
Annotation. ϕ – variance of the observed (exogenous) variable; ψ – residual variance of the latent (endogenous) variable.

The defined model proved significant ($\chi^2 = 82.64$; $df = 17$; $p < .001$). Both absolute and incremental fit indices met expectations and were high (GFI = .98; AGFI = .94; CFI = .95) (Hooper, Coughlan, Mullen, 2008; Kline, 2015). The relative fit index was satisfactory but lower than the previously mentioned parameters (TLI = .92). However, this is justified as TLI penalizes complex models (Hair et al., 2010). The noncentrality-based index for the error measure and its 95% confidence interval was low (RMSEA = .06, 95% CI [.03, .08]); the same was true of the absolute error index (SRMR = .03). Next, no standardized residual covariance was significant, thus indicating no substantial differences between the covariances based on the theoretical model and the observed ones. As it was possible to estimate the value of every parameter, the model was successfully identified. The fit indices were satisfactory if not exceptionally good (given the degrees of freedom). These results show that the theoretical model had good fit to the actual data from the SMAT-G measurement, thus supporting H2.

Lastly, group differences were tested based on participants' gender for every detailed and general SMAT-G result. Due to the normality of the distribution test results, Student's *t* was selected to examine between-group differences between the mean scale and the overall results. As shown in Table 3, there were no significant differences between men and women on any scale. This indicates that if SMAT-G were used, for example, to support employment decisions, it would not work in favor of either gender, therefore providing validity evidence based on test consequences. What should be emphasized is that this does not indicate the validity of SMAT-G as a tool for measuring cognitive ability, as the issue of gender differences in abilities testing is complex. It only implies that the consequence of using the test would not be a biased hiring decision.

Methods: Study 3 – Validity Examination

The last study aimed to confirm whether SMAT-G results are valid indicators of cognitive abilities. Two recognized, reliable, and valid tools were used as criteria: Cattell's Culture Fair Intelligence Test (CFT-3) (Matczak, Martowska, 2013) and the Word Understanding Test (TRSS) (Matczak, Jaworowska, Martowska, 2012). Each of them measures a different component of intelligence: fluid one (inductive reasoning) in the case of CFT-3 and crystallized one (lexical knowledge) in TRSS. Thus, significant relationships between the developed and criteria tests will provide evidence for the validity of overall SMAT-G. At the same time, higher correlations between specific SMAT-G tests (Gf_1 and Gc_3) and criterion tests that measure the same narrow abilities will indicate that the SMAT-G scales measure the target characteristics for which they were designed validly. The expected results may provide evidence based on test content, and convergent and construct validity based on APA and EFPA standards. Therefore, the following hypotheses were checked:

H7. CFT-3 results will correlate with SMAT-G overall results.

H8. CFT-3 results will correlate with Gf_1 scale results higher than with the other scales.

H9. TRS-S results will correlate with SMAT-G overall results.

H10. TRS-S results will correlate with Gc_3 scale results higher than with the other scales.

Sample Description

In this study, employees of private companies (n = 198) aged 33.20 years (SD = 11.43, with seven missing values) participated. On average, they had 9.26 years of professional experience (SD = 7.84). There were 80 men (42.3%) and 109 women (57.7%), and nine people did not provide this information.

Measurements

Besides the demographic survey, three cognitive tests were used.

SMAT-G. General and narrow abilities were measured with the final version of SMAT-G, with the overall score (GMA) calculated as the mean of its eight scales.

CFT-3. CFT-3 is a tool to assess fluid intelligence, primarily inductive reasoning. The test consists of four scales: series, classification, matrix, and topology. CFT-3 has two sets, A and B, but only A was used in the study (following the test guidelines). The test is time-limited. The overall result was used.

TRS-S. The TRS-S test is based on the synonyms method: participants must indicate synonyms for 32 keywords over a period of about 10 minutes. It measures crystallized abilities (general lexical knowledge, namely). The S version, which is intended for the general population, was used.

Procedure

The study was conducted among the employees of Polish companies (k = 19). Convenience sampling was used: team members contacted organizations' representatives via LinkedIn. After giving their initial agreement, the companies received information about the procedure (during online meetings with Q&A sessions) and the purpose of the study; subsequently, they collected information about the employees who agreed to participate. Researchers then sent the appropriate number of study materials to the companies (packaged in envelopes containing all three tools). Tests were distributed by the HR departments and then completed during group meetings. Participants first completed CFT-3 and then proceeded with SMATG and TRS-S. Team leaders or HR representatives administered the tests and held these meetings. The procedure took from 45 to about 70 minutes. All tests were returned directly to the researchers after being sealed. Employees who abandoned the procedure were asked to return the incomplete tests. The results of 31 people were excluded due to lack of data or withdrawal from the study.

Results: Study 3

Pearsons's r coefficients are listed in Table 4. CFT-3 presented a high correlation with the Gf_1 scale. Cattell's test results were also significantly correlated with the results of the Gf_2 scale, which measures serial reasoning ability (a type of induction, according to the CHC model). This is consistent with the theoretical basis of both tools. CFT-3 results also correlated significantly with the overall score from SMAT-G, thus supporting H7. The overall SMAT-G results correlated with the TRS-S score, supporting H9. The synonym test had a positive and significant correlation with the Gc_3 scale. The correlation was moderately strong; this can be explained by the different choice of words and the differences between the tests. The TRS-S score was also correlated with another scale measuring crystallized abilities, namely Gq_1. Altogether, these results are a positive indicator of the validity of SMAT-G in measuring cognitive ability. Next, tests for significant differences between independent correlation coefficients were conducted. Coefficients for each SMAT-G scale and CFT-3 results were tested against Gf_1 and CFT-3 Pearson's r.

For TRS-S, comparisons were performed against Gc_3 and TRS-S correlation coefficients. Results are presented in Table 4. In line with H8 and H10, a correlation of Gf_1 and CFT-3 was highest among all of the SMAT-G and Cattell's *r* coefficients, and Gc_3 and TRS-S correlation was significantly higher than any other of TRS-S's relations. These results support the validity of SMAT-G in GMA estimation and in the context of testing specific narrow abilities.

Table 4. SMAT-G, CFT-3, and TRS-S results' correlations

	r (CFT-3)	*z* (CFT-3)	*r* (TRS-S)	*z* (TRS-S)
Gf_1	.74***	–	.36**	4.47***
Gf_2	.45**	4.82***	.41**	4.00***
Gf_3	.38**	5.54***	.22*	6.19***
Grw_3	.44**	4.94***	.47**	3.15***
Grw_4	.46**	4.44***	.34*	4.35***
Gc_2	.55***	3.50***	.49**	2.89**
Gc_3	.35*	5.89***	.67***	–
Gq_1	.43**	5.19***	.47**	2.89**
GMA	.61***	2.45***	.53***	2.36***

Annotation. *r* – Pearson's *r* coefficients of SMAT-G scales correlations with CFT-3 or TRS; *z* – *z* coefficients for SMAT-G scales and CFT-3 independent correlation coefficients differences tests; * $p < .05$; ** $p < .01$; *** $p < .001$.

Discussion

Three studies were conducted, providing evidence for both the reliability and the validity of SMATG results as an estimation of cognitive abilities. Both the split-half coefficients and the good fit of the observed data to the theoretical model confirmed in Study 2 provide evidence for the reliability of the test results. The validity verification was comprehensive. Based on APA guidelines (1999), evidence based on test content (scale content analysis and the correlation matrix from Study 2), response processes (initial analysis and unprompted comments from participants), internal structure (CFA from Study 2), relation to other variables (correlations from Study 3), and the consequences of application (gender bias analysis in Study 2) were gathered. Similarly, both construct and criterion validity based on EFPA standards (2013) were proven. All of this suggests that SMAT-G can be used to test both general and specific cognitive abilities on the group level for research purposes.

The gathered data confirmed the CHC model as an accurate description of the structure of human intelligence. This was not the purpose of the conducted studies, as CHC has been confirmed numerous times (McGrew, 2005; Schneider, McGrew, 2012). What is essential, though, is that the SMAT-G results were in line with the CHC model; therefore, the test is a robust theory-based tool; its results can be utilized in the scientific inference process. Therefore, SMAT-G is suitable for at least two purposes. There is an ongoing debate about the specific validity theory and job performance (Kell, Lang, 2018). A tool for measuring numerous narrow abilities from various broader factors (suited for a work context) may help achieve further progress in this area. Second, SMAT-G proved to be a valuable element that can be included in the cross-battery assessment (XBA) procedure introduced by Woodcock (1990). The procedure mentioned above overcomes the difficulty of not having a single tool to comprehensively measure all abilities described in the CHC model. Woodcock proposed a series of rules for combining multiple tests to measure the complete range of cognitive characteristics as long as they are based on the shared theoretical ground. Since SMAT-G meets these assumptions and includes scales not found in other tests (such as Gq), it could be a valuable tool for practitioners and researchers interested in thoroughly assessing a wide range of abilities.

Limitations and Future Guidelines

Every effort has been made to ensure that the development and validation of SMAT-G were comprehensive and in accordance with psychometric standards. However, the tool requires further evaluation, so it is crucial to identify the limitations of its use. The choice of scales selected for the scope of the SMAT-G test may raise objections because it omits time-based tasks. This can lead to a limitation of the comprehensiveness of the evaluation of the abilities of a single subject. While acknowledging this limitation, one must bear in mind the purpose for which the SMAT-G was developed, i.e., to estimate general and specific cognitive abilities at the group level in order to be able to check hypotheses concerning the importance of these abilities in the work context. The test makes such estimations possible, as concluded from gathered evidence. It includes the two most important factors when it comes to the *g* loading, namely Gc and Gf. Thus, the estimated overall result of SMAT-G is a valid representation of GMA in a given group of subjects. The most severe limitation seems to be the issue of reliability and normalization; therefore, further research in this area is required before SMAT-G can be recommended for individual diagnosis. In addition, future development of the tool may include the introduction of subsequent scales corresponding to further broad abilities from the CHC model and confirmation of the relationship between its results and important real-life outcomes, such as professional success or academic performance.

Overall, the article presents the theoretical basis, i.e., the CHC model, and describes the development process and results of four validation studies that confirm that the SMAT-G test is a reliable and valid tool for measuring human intelligence at the group level.

References

Alfonso, V.C., Flanagan, D.P., & Radwan, S. (2005). The impact of the Cattell–Horn–Carroll theory on test development and interpretation of cognitive and academic abilities. In D.P. Flanagan, & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues. 2nd Ed.* (pp. 185–202). Guilford.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* American Educational Research Association.

Cambridge Brain Sciences. (2017). *Online Neurocognitive Tasks*, https://www.cambridgebrainsciences.com/science/tasks

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

Condon, D.M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence, 43*(1), 52–64, doi: 10.1016/j.intell.2014.01.004

Ebel, R.K., & Frisbie, D.A. (1986). *Essentials of educational measurement.* Prentice-Hall.

European Federation of Psychologists' Associations. (2013). *EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests, version 4.2.6*, http://www.efpa.eu/

Flanagan, D.P., Ortiz, S.O., & Alfonso, V.C. (2007). *Essentials of Cross Battery Assessment. 2nd Ed.* Willey.

Grobelny, J. (2018). Predictive Validity toward Job Performance of General and Specific Mental Abilities. A Validity Study across Different Occupational Groups. *Business and Management Studies, 4*(3), 1–12, doi: 10.11114/bms.v4i3.3297

Guttman, L., & Levy, S. (1991). Two structural laws for intelligence tests. *Intelligence, 15*(1), 79–103, doi: 10.1016/0160-2896(91)90023-7

Hair, J., Anderson, R., Tatham, R., & Black, W. (2010). *Multivariate Data Analysis. 7th Ed.* Prentice Hall.

Hooper, D., Coughlan, J., & Mullen, M.R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53–60.

Kaufman, A.S. (2009). *IQ Testing 101.* Springer Publishing.

Kell, H.J., & Lang, J.W.B. (2018). The Great Debate: General Ability and Specific Abilities in the Prediction of Important Outcomes. *Journal of Intelligence, 6*(3), 39, doi: 10.3390/jintelligence6030039

Kline, R.B. (2015). *Principles and practice of structural equation modeling. Third Edition.* Guilford Press.

Lang, J.W.B., & Kell, H.J. (2019). General Mental Ability and Specific Abilities: Their Relative Importance for Extrinsic Career Success. *Journal of Applied Psychology,* doi: 10.1037/apl0000472

Learning Express. (2005). *501 challenging logic and reasoning problems. 2nd Ed.* Learning Express.

Matczak, A., Jaworowska, A., & Martowska, K. (2012). *Test Rozumienia Słów – Wersja Standard i Wersja dla Zaawansowanych*. Pracownia Testów Psychologicznych PTP.

Matczak, A., & Martowska, K. (2013). *Neutralny Kulturowo Test Inteligencji – wersja 3 Raymonda B. Catella i Alberty K.S. Catell*. Pracownia Testów Psychologicnych PTP.

McGrew, K.S. (2005). The Cattell–Horn–Carroll Theory of Cognitive Abilities: Past, Present, and Future. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues. 2nd Ed.* (pp. 136–182). Guilford Press.

Nisbett, R. (2009). Intelligence and How to Get It: Why Schools and Cultures Count. In *Brock Education Journal*. W.W. Norton & Company, https://doi.org/10.26522/brocked.v19i2.138

Open Source Psychometrics Project. (2017a). *Full Scale IQ Test*, https://openpsychometrics.org/tests/FSIQ/

Open Source Psychometrics Project. (2017b). *Multifactor General Knowledge Test*, https://openpsychometrics.org/tests/MGKT2/

Richardson, K., & Norgate, S.H. (2015). Does IQ Really Predict Job Performance? *Applied Developmental Science, 19*(3), 153–169, doi: 10.1080/10888691.2014.983635

Rojon, C., McDowall, A., & Saunders, M.N.K. (2015). The Relationships Between Traditional Selection Assessments and Workplace Performance Criteria Specificity: A Comparative Meta-Analysis. *Human Performance, 28*(1), 1–25, doi: 10.1080/08959285.2014.974757

Rosseel, Y., & Jorgensen, T.D. (2020). *Package "lavaan". Latent Variable Analysis (wersja 0.6–6).* Comprehensive R Archive Network (CRAN).

Sackett, P.R., Zhang, C., Berry, C.M., & Lievens, F. (2021). Revisiting Meta-Analytic Estimates of Validity in Personnel Selection: Addressing Systematic Overcorrection for Restriction of Range. *Journal of Applied Psychology*, doi: 10.1037/APL0000994

Salgado, J.F. (2017). Using Ability Tests in Selection. In H.W. Goldstein, E.D. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention* (pp. 115–150). Wiley-Blackwell.

Salgado, J.F., Moscoso, S., De Fruyt, F., Anderson, N., Bertua, C., & Rolland, J.P. (2003). A Meta-Analytic Study of General Mental Ability Validity for Different Occupations in the European Community. *Journal of Applied Psychology, 88*(6), 1068–1081, doi: 10.1037/0021-9010.88.6.1068

Schmidt, F.L., & Hunter, J.E. (1993). Tacit Knowledge, Practical Intelligence, General Mental Ability, and Job Knowledge. *Current Directions in Psychological Science, 2*(1), 8–9, doi: 10.1111/1467-8721.ep10770456

Schneider, W.J., & McGrew, K.S. (2012). The Cattell–Horn–Carroll Model of Intelligence. In D.P. Flanagan, & P. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues. 3rd Ed.* (pp. 99–144). Guilford Press.

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. 4th Ed.* (pp. 73–163). Guilford Press.

Schneider, W.J., & Newman, D.A. (2015). Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. *Human Resource Management Review, 25*(1), 12–27, doi: 10.1016/j.hrmr.2014.09.004

Sourceforge. (2012). *PEBL Psychological Test Battery*, http://pebl.sourceforge.net/battery.html

Stankov, L. (1997). *The Gf/Gc Quickie Test Battery. Unpublished test battery from the School of Psychology*. University of Sydney.

Sternberg, R.J., Kaufman, J.C., & Grigorenko, E.L. (2008). *Applied Intelligence*. Cambridge University Press.

Viswesvaran, C., & Ones, D.S. (2000). Perspectives on Models of Job Performance. *International Journal of Selection and Assessment, 8*(4), 216–226, doi: 10.1111/1468-2389.00151

Wagner, R.K., & Sternberg, R.J. (1987). Tacit Knowledge in Managerial Success. *Journal of Business and Psychology, 1*(4), 301–312, doi: 10.2307/25092106

Woodcock, R.W. (1990). Theoretical Foundations of the Wj-R Measures of Cognitive Ability. *Journal of Psychoeducational Assessment, 8*(3), 231–258, doi: 10.1177/073428299000800303

OPRACOWANIE I WALIDACJA TESTU ZDOLNOŚCI POZNAWCZYCH SMAT-G

Streszczenie. Artykuł opisuje proces opracowania nowego narzędzia do pomiaru zdolności poznawczych: SMAT-G. Test oparty jest na modelu inteligencji Cattella–Horna–Carrolla (tj. trzystopniowej strukturze zdolności poznawczych). Opisane zostały kolejne kroki wyboru skali (w tym mierzących zdolności płynne i skrystalizowane, czytanie i rozumienie oraz wiedzę ilościową) oraz proces opracowania pozycji testowych. Przeprowadzono trzy badania walidacyjne i przedstawiono ich wyniki. Przeanalizowane zostały procesy udzielania odpowiedzi na pozycje testowe, wskaźniki trudności i mocy dyskryminacyjnej, rzetelność połówkowa, struktura czynnikowa i możliwe konsekwencje stosowania testu. Wyniki SMAT-G odpowiadają modelowi teoretycznemu i wykazują istotne korelacje z uznanymi testami zdolności poznawczych. Przeprowadzone badania potwierdzają rzetelność i trafność SMAT-G.

Słowa kluczowe: test zdolności poznawczych, specyficzne zdolności poznawcze, ogólne zdolności poznawcze, inteligencja płynna, inteligencja skrystalizowana, testowanie

Receipt Date: 8th November 2022
Receipt Date after correction: 12th March 2023
Print Acceptance Date: 16th March 2023