AUTHOR	Haataja Eeva S.H., Tolvanen Asko, Vilppu Henna, Kallio Manne, Peltonen Jouni, Metsäpelto Riitta-Leena
TITLE	Measuring higher-order cognitive skills with multiple choice questions –potentials and pitfalls of Finnish teacher education entrance
YEAR	2022
DOI	https://doi.org/10.1016/j.tate.2022.103943
VERSION	Final draft
CITATION	Haataja Eeva S.H., Tolvanen Asko, Vilppu Henna, Kallio Manne, Peltonen Jouni, Metsäpelto Riitta-Leena (2022). Measuring higher-order cognitive skills with multiple choice questions –potentials and pitfalls of Finnish teacher education entrance. Teaching and Teacher Education https://doi.org/10.1016/j.tate.2022.103943

Abstract

This mixed methods study examines the structure of the multiple-choice exam for student selection in Finnish teacher education. Through qualitative content analysis, we categorized multiple-choice questions into items that assessed lower- and higher-order cognitive processes based on the Revised Bloom's Taxonomy. Exploratory and confirmatory factor analyses yielded four factors that represented lower- and higher-order cognitive processing skills and comprehension of empirical and theoretical items. These were associated with matriculation examination grades, especially with the average grade and the mother tongue grade. When developing future multiple-choice exams for admissions, we recommend emphasizing higher-order processing skills and the role of source materials.

Keywords: higher-order processing skills, multiple-choice questions, student selection, teacher education, Revised Bloom's Taxonomy

1. Measuring Higher-Order Cognitive Skills with Multiple-Choice Questions: Potential and Pitfalls of Finnish Teacher Education Entrance

Articulating and implementing high-standard teaching and teacher education have been attempted around the world in recent years, with the Finnish educational system often being regarded as an example in this respect (Darling-Hammond, 2017; Malinen et al., 2012). One notable feature of Finnish initial teacher education (ITE) is the selection of students at the entry phase with the aim of identifying applicants who have strong potential to develop into teachers. Student selection for ITE should not be based on the attributes of a good teacher alone but on empirical evidence of effective and reliable selection methods (Bardach et al., 2022), which must therefore be topics of active research. This study examines Finnish student selection for ITE programs with the aim of analyzing a multiple-choice exam measuring applicants' higher-order processing skills, the so-called VAKAVA exam (VAKAVA is the Finnish acronym for the National Educational Selection Cooperation Project; Malinen et al., 2012).

The work of teachers has a strong impact on the learning, motivation, and well-being of children and young people; therefore, decisions regarding admitted and failed applicants for ITE bear societal impacts (Klassen & Kim, 2019). In recent years, Finnish universities have undergone significant reforms in the methods and processes of student selection. For ITE, this means emphasizing the role of matriculation examination (ME; exit examination of secondary education that provides eligibility for tertiary education) and replacing admission tests that require laborious advanced preparation with tests that provide all source materials on the test site. A carefully planned and implemented student selection process helps ensure the optimal use of resources in the educational system (Kuncel et al., 2001) through improvements in teacher

efficacy and professional well-being (Bardach et al., 2022). However, unlike many other fields of the teaching profession, research on student selection has long been scarce and has only recently begun to increase (Klassen & Kim, 2019; Metsäpelto et al., 2022). Reflections on the objectives and utility of admission tests are required, especially during times of reform in admission policy (Parvaneh, 2020; Thomson et al., 2011).

In Finland, student selection for ITE evaluates applicants' cognitive and non-cognitive skills in two sequential phases (see Figure 1 and section 4 for more information on Finnish selection for ITE). This study focuses on the first, cognitive evaluation phase and particularly on the VAKAVA exam, a multiple-choice exam to assess applicants' cognitive processing skills. Higher-order cognitive skills, such as abstract thinking, comprehension of complex ideas, and fast learning, are highly relevant for teaching because of the complexity of the profession. Learning general cognitive processing skills is not at the heart of Finnish ITE, which underlines the need for students to already have these abilities upon entrance to ITE (Metsäpelto et al., 2021). Although success in cognitive admission tests does not always predict teacher effectiveness in working life (Bardach & Klassen, 2020), it has strong relevance for success in ITE studies. Admission tests that simulate academic studies are found to be valid predictors of future academic performance, and success in these tests is an indication of a person's general ability to perform in exams (Niessen et al., 2018).

As this is the first study to investigate the structure of the VAKAVA exam, we chose to use a mixed methods approach. The specific goal of this research is to utilize the Revised Bloom's Taxonomy (RBT; Krathwohl & Andersson, 2001), which is a widely used tool to analyze and categorize the level of cognitive processing in testing and exams (Newton & Martin, 2013), and

qualitatively analyze the cognitive processing skills required to succeed in the VAKAVA multiple-choice exam. We also aim to specify the cognitive processing skills in the VAKAVA exam *quantitatively* using advanced statistical approaches and to examine the linkages between the scores earned in the exam and prior academic achievement. This study provides new information about the multiple-choice format in student selection for ITE, which can also be applied more broadly when designing and improving multiple-choice questions in educational assessment.

2. Cognitive Processing Skills in Teacher Education

General cognitive abilities create the foundation for processing and acquiring work-related knowledge (Kuncel et al., 2001). Therefore, higher-order cognitive processing skills—for instance, the skills to analyze, reason, and solve problems and to comprehend and apply complex ideas—are essential for both successful studying to become a teacher and practicing in the teaching profession (Metsäpelto et al., 2021). The RBT is a theory on cognitive processing skills (Krathwohl & Andersson, 2001). It is a robust theory that can be applied in diverse contexts (Campbell et al., 2019); for instance, to analyze the cognitive processing skills required in admission tests.

The VAKAVA exam follows a multiple-choice format and requires applicants to use the source materials presented to them on-site in order to respond to questions. As these source materials include scholarly papers from the field of education, the participants need not only cognitive processing skills but also scientific reading skills to be able to extract correct knowledge from the materials and succeed in the exam.

2.1 Lower- and Higher-Order Cognitive Processing Skills

The distinction between lower- and higher-order cognitive processing skills ensues from Bloom's Taxonomy (BT; Bloom, 1956). The RBT (Krathwohl & Anderson, 2001) is a two-dimensional construct of the objectives of a curriculum and for instruction that has been widely used to classify educational and curricular goals and to categorize learning outcomes across many disciplines (Fuller et al.,2007; Hanna, 2007). The RBT has also been widely used to examine the cognitive demand of exams and testing (Newton & Martin, 2013). The RBT presents the continuum of cognitive processes and the knowledge expected to be acquired (Krathwohl & Anderson, 2001).

In the original BT, as well as in the RBT, the levels of cognitive processes are hierarchic, meaning that mastery of lower levels is required to acquire mastery of higher levels (Zheng et al., 2008). In professional work, both higher- and lower-order thinking skills are needed, and recent research examines these skills as qualitative differences in students' abilities that should be assessed separately to avoid bias in interpreting them (DiDonato-Barnes et al., 2014; Jansen & Möller, 2022). Additionally, the level of the objective in an exam does not directly indicate the importance (Jensen et al., 2014) or difficulty (Thompson et al., 2013) of the performance. Small pieces of information can be very significant in professional work or cumbersome to recall.

The dimension of cognitive processes includes six levels. The first two levels, Remembering and Understanding, are *lower-order cognitive processes*. *Remembering* refers to a person's ability to retrieve information from memory that is relevant for solving a task. *Understanding* refers to constructing the meaning of the source materials by interpreting, exemplifying, classifying, summarizing, inferring, comparing, or explaining information. In the

original BT, the levels of Remembering and Understanding were called Knowledge and Comprehension, respectively, and were referred to as static objectives of teaching, whereas in the RBT, they are more oriented toward assessing dynamic processes of thinking (Krathwohl & Anderson, 2001).

The remaining four levels, Applying, Analyzing, Evaluating, and Creating, represent *higher-order cognitive processes*. Applying means using a procedure, either introduced or unfamiliar to the participants, in a new context through processes of execution or implementation. Analyzing refers to differentiating or organizing parts of information and reflecting on the relations between these parts. These two levels are the highest forms of cognitive processing that can be assessed with multiple-choice tasks. The highest levels of the taxonomy, Evaluating and Creating, cannot be assessed using multiple-choice items and were therefore not applied in this study (Krathwohl & Anderson, 2001).

2.2 Scientific Reading Skills

In the VAKAVA exam, applicants read scholarly papers in the field of education on the test site and respond to multiple-choice questions based on such papers. This design resembles other large-scale exams, such as IELTS and TOEFL, which are both English language proficiency exams; they all include reading comprehension of academic texts on-site, combined with multiple-choice questions (Baghaei et al., 2020).

While learning to read scientific texts can be considered a crucial element in teachers' professional development, this might be new to ITE applicants. The argumentative style and technical language of scientific literature have been shown to cause difficulties for novice student

readers (van Lacum et al., 2012; Yarden, 2009). Generally, sentence length and familiarity with the words in a text, as well as the number, coherence, and structure of ideas expressed, contribute to reading difficulties (Kintsch, 2004). Some task-related factors may also affect the difficulty of the task as a function of its cognitive demands, such as questions requiring the reader to gather multiple pieces of information across texts (Organization for Economic Co-operation and Development, 2021). General reading fluency, defined as an individual's ability to read texts quickly and accurately (Kuhn & Stahl, 2003), contributes to reading more complex texts. Fluent readers have more cognitive resources left for higher-level comprehension processes, such as reading strategies and inferences (Walczyk et al., 2004).

The use of demanding scientific papers in an admission test exam with applicants having little experience in scientific reading is believed to increase the discriminative power of the exam and help identify those applicants who already have strong scientific reading skills acquired through secondary education. It is noteworthy that prior research on materials selected as sources of information in a multiple-choice question format has been scarce. However, the quality of source materials influences the demandingness of the exam, that is, the level of scientific reading and cognitive processing skills required to respond to the items (e.g., Haladyna, 2004). One form of scientific paper, an empirical paper, presents research based on concrete observations and empirical data, whereas a theoretical paper builds on theories and concepts that are used to generate novel insights (Jaakkola, 2020).

3. Assessing the Level of Cognitive Processing Skills Using Multiple-Choice Question Exams

Previous research indicates that both lower- and higher-order cognitive processes, as described by BT, can be assessed using multiple-choice questions (Case & Swanson, 1998; Zaidi et al., 2018; Zheng et al., 2008). Lower-order multiple-choice questions require the respondent to recall or comprehend information, and these tend to be easier for applicants (Zaidi et al., 2016). Higher-order multiple-choice questions require deeper processing, such as applying information in a new situation, drawing conclusions, discerning relevant from irrelevant information, and identifying relationships between methods, concepts, principles, and theories (Jensen et al., 2014; Krathwohl & Anderson, 2001; Zaidi et al., 2018). Success in both low- and high-order tests is related to success in later exams during a learner's studies (Zaidi et al., 2016).

Multiple-choice questions are a reliable and cost-efficient method of assessing knowledge in a particular topic but have been criticized for measuring the recognition of trivial knowledge and factual recall rather than higher-order cognitive processing (Case & Swanson, 2001; Van der Vleuten, 1996). Many studies report that the cognitive levels of multiple-choice exams have been modest. For example, Masters et al. (2001) found that two thirds (69%) of the items in test banks in nursing education were written at a lower-order cognitive level. Similarly, some other studies have reported that lower-order cognitive levels of BT comprise over 50% (Palmer & Devitt, 2007; Zheng et al., 2008) to up to 90% of test items (Tarrant et al., 2006). Notably, the criticism of the low cognitive level of the multiple-choice question format is primarily directed at test writers, not the multiple-choice format *per se* (Haladyna, 2004).

Therefore, more effort should be directed at designing multiple-choice questions in a way that includes a balanced combination of lower- and higher-order thinking, which is a challenging

goal (Zheng et al., 2008). As noted above, the materials selected as sources of information affect the level of cognitive processing that can be assessed through them (Krathwohl & Anderson, 2001). Students who struggle with answering multiple-choice questions rate the items as requiring higher-order processing according to BT, even if teachers have purposed them as lower-order items (Stringer et al., 2021). Additionally, students' situational learning approach has an impact on the depth of the cognitive processes on BT: does the situation support the students' willingness to obtain deep understanding or just pass the test to achieve other goals, such as admission to university studies (Zaidi et al., 2018)? A study that categorized undergraduate assessment tasks using BT found that success in multiple-choice tasks of the Applying level was predicted by a deep approach to learning (Newton & Martin, 2013).

The method used to determine the cognitive level of multiple-choice items has typically been qualitative content analysis, in which the content of the items is categorized using a predefined coding scheme. Various studies have underlined that this is a complex task (e.g., Baghaei et al., 2020). Therefore, reaching an adequate level of interrater reliability is highly important. Conducting the coding separately by two researchers, comparing the categories, and finding consensus through discussions are widely used methods to ensure the reliability of this qualitative analysis (Neiro & Johansson, 2020; Thompson et al., 2013; Zheng et al., 2008).

The difficulty of reaching a reliable categorization partly stems from the obscurity of the categories that can be seen to overlap (Thompson et al., 2013). Based on empirical findings and theoretical principles, many studies suggest collapsing neighboring categories into two main groups: lower- and higher-order processing (DiDonato-Barnes et al., 2014; Jensen et al., 2014).

This dichotomic examination of the levels of cognitive processing also diminishes issues related to the original epistemic assumptions of BT (DiDonato-Barnes et al., 2014).

Few studies have attempted to examine the cognitive levels of multiple-choice exams by investigating their internal factor structures using advanced quantitative analyses. This may be because multiple-choice exams often do not have hypothesized dimensionality defined *a priori*; therefore, successful attempts to empirically define factor structures could benefit the field. In the present study, we aimed to address this question and used quantitative analysis of factor structures (i.e., exploratory factor analysis [EFA] and confirmatory factor analysis [CFA]) to identify the cognitive processing skills required to respond to multiple-choice questions and to specify the relationships between them.

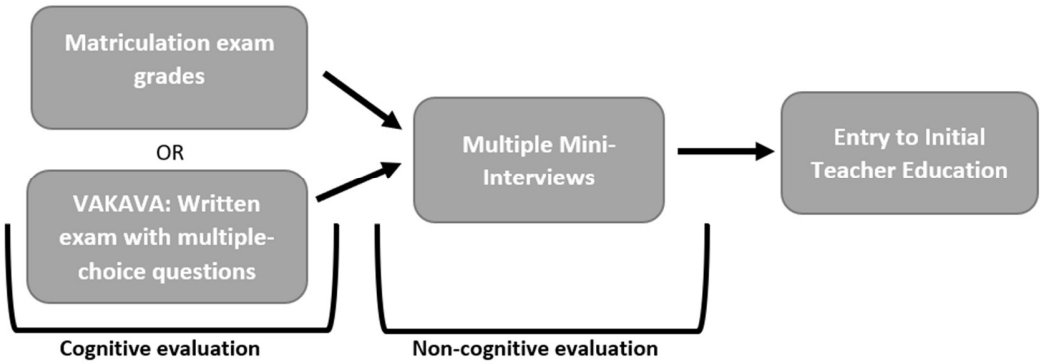
4. The Context: Finnish Student Selection for ITE

Finland is among those countries that highly appreciate the teaching profession; unlike many countries, it strongly controls entry to ITE (Darling-Hammond, 2017; Hammerness et al., 2017; Ingvarson, 2013). Finnish ITE programs consist of a three-year bachelor's degree followed by a two-year master's degree, including studies in pedagogy, communications, and research methods. Class and special education teacher students graduate at the master's level, whereas early childhood education teacher students receive their teacher's degrees at the bachelor's level but may continue to master studies if they aim for leadership positions in the educational field. The popularity of ITE programs in Finland, with significantly more applicants than accepted students (Hammerness et al., 2017; Malinen et al., 2012), makes the admission tests high-stakes

test situations for applicants (Cross & O'Loughlin, 2013). The Finnish selection process for ITE programs combines assessment of applicants' cognitive and non-cognitive attributes (Figure 1).

Figure 1

Structure of the Admissions Process for ITE in Finland



The first phase of selection consists of cognitive evaluation, which includes matriculation examination scores or the VAKAVA exam. The top applicants with the highest ME scores proceed directly to the second phase of selection, which evaluates non-cognitive attributes by means of multiple mini-interviews. Applicants with lower ME scores can seek entrance to the second phase by taking the VAKAVA exam. The purpose of having the VAKAVA exam as an option for those with ME scores is to increase equity between applicants. It offers an opportunity for applicants whose ME grades do not for some reason represent their level of cognitive skills; such reasons may include personal struggles during high school (Hammerness et al., 2017). The final selection decision is made based on multiple mini-interviews.

Students complete the ME to graduate from general upper secondary school, and it is the only national high-stakes test in the Finnish school system (Kupiainen et al., 2016). The exam consists of tests in at least five subjects based on the student's choice; however, the test in

mother tongue and literature, usually Finnish, is mandatory. The other tests are in mathematics, foreign languages (separately in each language), and several other subjects in the fields of humanities and natural sciences. Test scores are scaled following the normal distribution across seven grades. Five percent of the candidates in each test will receive the highest grade, and so on, following the distribution of 5% (the highest grade), 15%, 20%, 20%, 20%, 15%, and 5% (meaning failing the test). Approximately 30,000 students pass the examination every year (Matriculation Examination Board, 2022).

The combination of both cognitive and non-cognitive attributes in Finnish student selection reflects the need for versatile competences in the teacher's work. Strong cognitive abilities are believed to help student teachers acquire, process, and construct knowledge in ITE and at work, while non-cognitive attributes help them, for instance, build positive student–teacher relationships, adaptively regulate emotions, stay motivated in the teacher profession, and create learning environments that foster pupils' learning, motivation, and well-being (Metsäpelto et al., 2021; Moè & Katz, 2020; Taxer & Gross, 2018). This study focuses on investigating the VAKAVA exam as a measure of cognitive processing skills.

Both VAKAVA exam scores and ME grades are indicators of a person's cognitive abilities because they both require cognitive processing of knowledge. The most significant difference between them is that ME grades measure learning and knowledge in a variety of disciplines accumulated across three years of upper secondary school, whereas the VAKAVA exam, in its current form, requires fast cognitive processing of information. The key features of the latter, namely processing speed, memory, and reasoning, are linked with general intelligence (Kail, 2000). Prior research on ME grades and VAKAVA scores has reported correlations ranging from

around .40 (Metsäpelto et al., 2019; Utriainen et al., 2012) to .60 (Räihä, 2010), although these findings were based on a prior version of the VAKAVA exam that required applicants to prepare for the test by learning scholarly papers in advance.

In the present study, we investigate the association between ME grades and VAKAVA exam scores to provide new knowledge about the relationship between these two types of educational assessment. We selected ME grades based on their relevance to the educational field or to the VAKAVA exam. The VAKAVA exam requires (a) reading comprehension, so mother tongue is included in this study, and (b) reasoning, so basic and advanced mathematics are included. We also selected psychology, social studies, and health education because psychological and societal questions and students' well-being are central areas of content in the Finnish ITE (e.g. University of Jyväskylä, 2020).

5. Research Questions

This mixed methods study investigates the structure of a multiple-choice question exam, the VAKAVA, which was developed to select students for ITE, and associations between VAKAVA exam scores and the ME grades. The study combines the qualitative and quantitative approaches, particularly in the data analysis and data interpretation phases. The qualitative and quantitative phases occur sequentially and are given approximately equal weight. We will address the following research questions (RQs):

1. What cognitive processing skills are represented in the items of the VAKAVA exam when its content is analyzed using the Revised Bloom's Taxonomy as a conceptual framework?

2. What kind of factor structure does the statistical analysis of the VAKAVA exam produce, and can the factors be interpreted by the Revised Bloom's Taxonomy?
3. What kind of an association can be found between the VAKAVA exam and matriculation exam grades?

6. Methods

6.1 Participants

The participants in this study were the entire population of applicants (N=6077) seeking admission to Finnish ITE programs (i.e., class, special education, early childhood education, craft, and subject teacher programs) in eight universities in 2021. The data were drawn from the national register for study programs leading to a degree, maintained by the Finnish National Agency for Education. In accordance with the European General Data Protection Regulation, the participants were informed about the processing of their personal data and given the opportunity to refuse participation. After removing the data of three applicants requesting non-participation in the study, we obtained a sample size of 6074. We used this sample to answer the first RQ. All the data were handled without personal information and stored confidentially.

Further analyses to respond to the second and third RQs were conducted with a subsample that included applicants for the classroom (grades 1–6; *n*=2621), early childhood education (*n*=812), and special education programs (*n*=562) who had graduated from Finnish high schools in 1990 or later (*N*=3994). In terms of the number of applicants and the vacancies available, these were the three largest ITE programs.

6.2 Measures

In the VAKAVA exam, applicants review scholarly papers in the field of education on-site and, using their own laptops, respond to multiple-choice questions, which include a lead-in question (stem), followed by one correct and one or more incorrect options. In 2021, the source materials consisted of two research papers published in Finnish educational journals in the Finnish language—the study by Peltola et al. (2020) and that by Raatikainen (2015). Peltola et al.'s (2020) paper was a qualitative empirical study on the aspects that worry student welfare professionals. Raatikainen's (2015) theoretical paper reflected on causal explanation in the social sciences. Applicants have three hours to complete the exam.

The exam consisted of 12 lead-in questions, each of which included several items. Altogether, the exam included 106 multiple-choice items. Each correct answer was graded with +1 point, leaving the item empty with 0 points, and a wrong answer with −0.33 to −1 point, depending on the number of false choices in the item. The number of true/false items was 87, and that of single best answer items was 19. For the statistical analyses, we coded the items as categorical variables with values of 1 for a correct answer, 0 for empty, and −1 for a wrong answer. The fourth and fifth authors of this study were members of the committee that developed the 2021 exam.

Among the ME data, we used the grades in mother tongue, basic and advanced mathematics, and three subjects representing the social sciences (psychology, health education, and social studies). We also used the ME total grade, which was calculated as the average of all subject scores the applicant obtained. The grades were converted to numeric values as follows (from best to worst): laudatur = 7, eximia cum laude approbator = 6, magna cum laude

approbator = 5, cum laude approbator = 4, lubenter approbator = 3, approbator = 2, and improbatur = 0, failed.

6.3 Qualitative Content Analysis

The first and third authors conducted the qualitative categorization of the VAKAVA exam items according to the RBT (RQ1) in three phases. First, we acquainted ourselves with the literature on using the RBT on multiple-choice items in high-stakes contexts to create a coding rubric (Table 1). In line with previous research (e.g., Zaidi et al., 2018), we chose to simplify the coding rubric to ensure the reliability and unambiguity of the coding. The categories of Remembering and Understanding represented the lower-order processes, whereas the categories of Applying and Analyzing represented the higher-order processes (cf. Jensen et al., 2014).

Table 1

The Qualitative Coding Rubric, Based on Jensen et al. (2014) and Krathwohl and Anderson (2001)

RBT code	Criteria	Example item
Lower-order cognitive processes	Term or definition can be found directly in the text (Remembering) or is reworded (Understanding)	*3_14 Randomizing the participants (e.g., students) into two groups guarantees that the researchers will find a causal relation.* [Said in the source material in almost the same way.]
Higher-order cognitive processes	The information is to be applied in a new context (Applying); information from different sources is to be combined, or parts of information are to be distinguished (Analyzing)	*3_36 Psychology from the standpoint of the subject is a special case of research that uses the covering law model for explaining human behavior.* [Requires understanding philosophical concepts of two papers and comparing them with each other.]

When coding the items, we considered each item as a whole, focusing on the formulation and wording of the stem, the information provided in the stem, the content of the question, and the response options. In most cases, the stem consisted of a general question or instruction, followed by multiple relatively similar items. In these cases, the entire task was coded into the same cognitive level category. As an exception, one task included 37 independent true/false items representing different levels of cognitive processes, and they were all coded separately. We followed the coding guidelines of Krathwohl and Andersson (2001), who advised interpreting the objectives of the exam "in relation to the meaning of the objective, the purpose of the instructional activities, and the aim of the assessments" (p. 97) to understand the statements and the cognitive activities in the items.

Therefore, we considered the purpose of the activity (Krathwohl & Andersson, 2001) to influence the applicants' cognitive processing when responding to multiple-choice questions (e.g., Zaidi et al., 2018). Given the high-stakes admission context, we assumed that the participants' approaches to the exam would be to answer the questions effectively and rapidly, rather than to learn deeply about the source materials. Therefore, we assumed that if the items could be answered without higher-order processing skills, the participants did not choose to use such skills, regardless of the actual complexity of the concepts of the task. These items were coded to the category of lower-order cognitive processes. Notably, the applicants had the source materials available throughout the exam, so whether they answered these items by recalling or re-reading the texts remained unclear. However, it is likely that the strict time limit (3 hours) forced them to use all available resources, including retrieval from memory, when responding.

After agreeing on the coding scheme, the first and third authors separately coded a sample of the items (*n*=23, 22%) and compared the coding afterward. In a discussion, a consensus on the coding of each item was reached. The researchers then continued coding the rest of the items independently and discussed the coding once more until complete agreement was reached.

6.4 Quantitative Analyses

To analyze the factor structure (RQ2) of the VAKAVA exam with EFA, we used Mplus (Muthén & Muthén, 1998–2017) to find the number of factors and items that significantly loaded to the factors. We conducted EFA with a weighted least square mean and variance adjusted estimator (WLSMV). The identified structure was further modified using CFA and modification indices. We allowed between-item correlations within the factors but did not allow cross-loadings (i.e., item loading into two factors). Finally, with structural equation modeling, we examined the relations of latent factors based on the CFA of the applicants' ME grades (RQ3).

The goodness-of-fit of the models was evaluated using (a) the χ2 test (nonsignificant *p*-values indicate a good fit), (b) the comparative fit index (CFI; values above .90 indicate an acceptable fit, and values close to 0.95 or above indicate a good fit), (c) the Tucker–Lewis Index (TLI; values above .90 indicate an acceptable fit, and values close to 0.95 or above indicate a good fit), (d) the root mean square error of approximation (RMSEA; values of 0.06 or less indicate a good fit), and (e) the standardized root mean residual (SRMR; values equal to or below 0.08 indicate a good fit) (Hu & Bentler, 1995). However, we ignored the significant *p*-value in the χ2 test, as our large sample size affected the reliability of that indicator (Hoe, 2008). Composite

reliability (CR) was used to evaluate the reliability of the factors (values equal to .70 or above indicate satisfactory reliability) (Hair et al., 2011).

After obtaining an acceptable model fit, we qualitatively examined the items of the factors from a theoretical and an empirical perspective to name such factors. For example, we observed that the items in one factor were based on the theoretical paper or on comparisons between the papers (Theoretical factor), whereas another factor concerned items on the empirical paper or the empirical section of the theoretical paper (Empirical factor). The third factor included only items that were coded as higher order (Higher Cognitive factor), and the fourth factor included items that were coded as lower order (Lower Cognitive factor) in the qualitative coding.

7. Findings

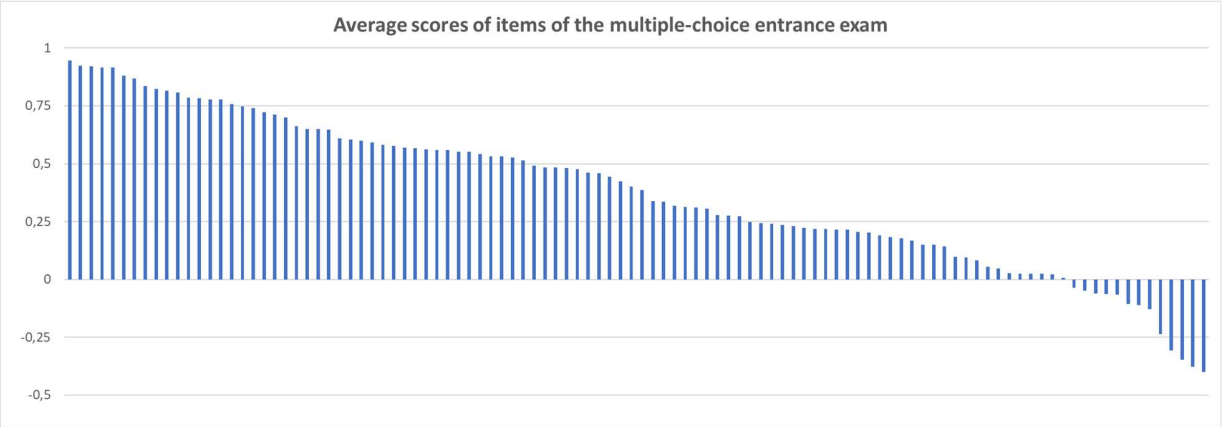
7.1. Descriptive Statistics

A preliminary analysis of the multiple-choice exam showed that it produced a wide dispersion of scores among the applicants. The applicants' total scores ranged from −31.65 to 100.00, with a mean score of 39.7 (*SD*=19.7). We also found that the frequency of negative and positive scores, indicating false and correct responses, respectively, and that of non-responses varied greatly across items. The frequency of false responses ranged from 56 to 3765 across items; that of correct responses ranged from 1095 to 5800; and that of non-responses ranged from 94 to 2902, suggesting large differences in item difficulty. This observation was supported by the findings presented in Figure 2, which illustrates the average scores obtained by the

applicants across all 106 items. The average scores of individual items ranged from 0.95 to −0.40 (*M*=0.37, *SD*=0.33), and 67 items (63%) lay within one standard deviation of the mean.

Figure 2

Distribution of the Average Scores for the Items Among the Applicants of the Finnish Entrance Exam for Teacher Education in 2021



The descriptive analysis of the ME scores showed that the applicants completed varying combinations of matriculation subjects (Table 2). Almost all applicants had taken a compulsory exam in the subject of mother tongue, whereas attending the ME in advanced mathematics and social studies was clearly less frequent. On a scale from 2 (lowest) to 7 (highest), the applicants' mean scores in psychology and health education were the highest, whereas those in advanced mathematics were the lowest.

Table 2

Descriptive statistics on the matriculation exam grades of the applicants

	N	*Mean*	*SD*
ME average grade	3994	3.77	1.94
Mother tongue	3891	3.99	1.09
Mathematics (advanced)	862	3.37	1.13
Mathematics (basic)	2208	3.68	1.23
Psychology	1572	4.10	1.15
Health education	1791	4.14	1.32
Social studies	781	3.86	1.19

7.2 Qualitative Content Analysis Identifying Items Representing Higher- and Lower-Order Cognitive Processing Skills

The findings of the content analysis according to the RBT showed that 64 items (60.38%) represented lower-order cognitive processes, that is, Remembering and Understanding. The correct answers to these items could be directly found in the source materials, so achieving the correct answer required only understanding and recalling explicit parts of the written source material. The following excerpt illustrates an item (1_5) that was coded into the lower-order cognitive category. It was the fifth item in the first task of the test and required the applicant to determine, *"Are the following statements right or wrong?"*

> *Worrying about one's own well-being is the least of the worries of the professionals of student welfare services.*

Sometimes, the concepts to which the lower-order items related were broad and complex in nature (e.g., *causal relation*), but answering the item correctly did not require absorbing the concepts but only finding the correct phrase in the text. The next example illustrates an item of this type (3_15).

> *Which of the following items are correct according to the papers?*
>
> *Correlational relation can sometimes be causal relation.*

The findings further showed that 42 (39.62%) of the items required higher-order cognitive processing. These items required applying information from the written source materials to a new context, such as an example of a practical situation in a school context (item 13_3):

Here are statements on students' undesired behavior in lessons. Associate each statement to the model or theory which it represents the most clearly in its means of explaining.

A. Hempel's model, B. Contrafactual theory, C. Interventionistic theory, D. None of the aforementioned

If a teacher notices that the students disagree on the desirable behavior with her, she should guide the students to negotiations on shared values.

This category also included items that required combining information from both research papers in the source materials. The knowledge in these items concerned larger concepts (such as the theories presented in the papers) and involved understanding the relations between them (items 3_31 and 3_36). The additions in brackets were not included in the original items.

Which of the following items is correct according to the papers?

Peltola et al.'s study [the empirical paper] *adopts the approach of positivistic research tradition* [the theoretical paper].

The Standpoint of the Subject [the empirical paper] *is a special example of research that explains human behavior with covering law model* [the theoretical paper].

All items in the exam could be categorized into either category. The mean of the applicants' scores in items representing lower-order cognitive processes was 0.45 (*SD*=0.44), whereas that of the applicants' scores in items representing higher-order cognitive processes was 0.26 (*SD*=0.40). The comparison showed a statistically significant difference between the mean scores ($t(104)=3.08$, $p=.003$), with the items requiring lower-order cognitive skills being easier.

The items also differed in terms of whether they were based on theoretical or empirical content. The items from the theoretical paper mainly included theoretical content. However, the theoretical paper also presented some examples of empirical studies, and the items based on these sections were about empirical content. Conversely, most of the items in the empirical paper were empirical by content, but there were some that concerned the research paradigm of the empirical paper, and these were categorized as theoretical. Some items also required a comparison between the empirical and theoretical papers, and these were categorized as comparative.

7.3 The Factor Structure of the Multiple-Choice Exam: Exploratory and Confirmatory Factor Analyses

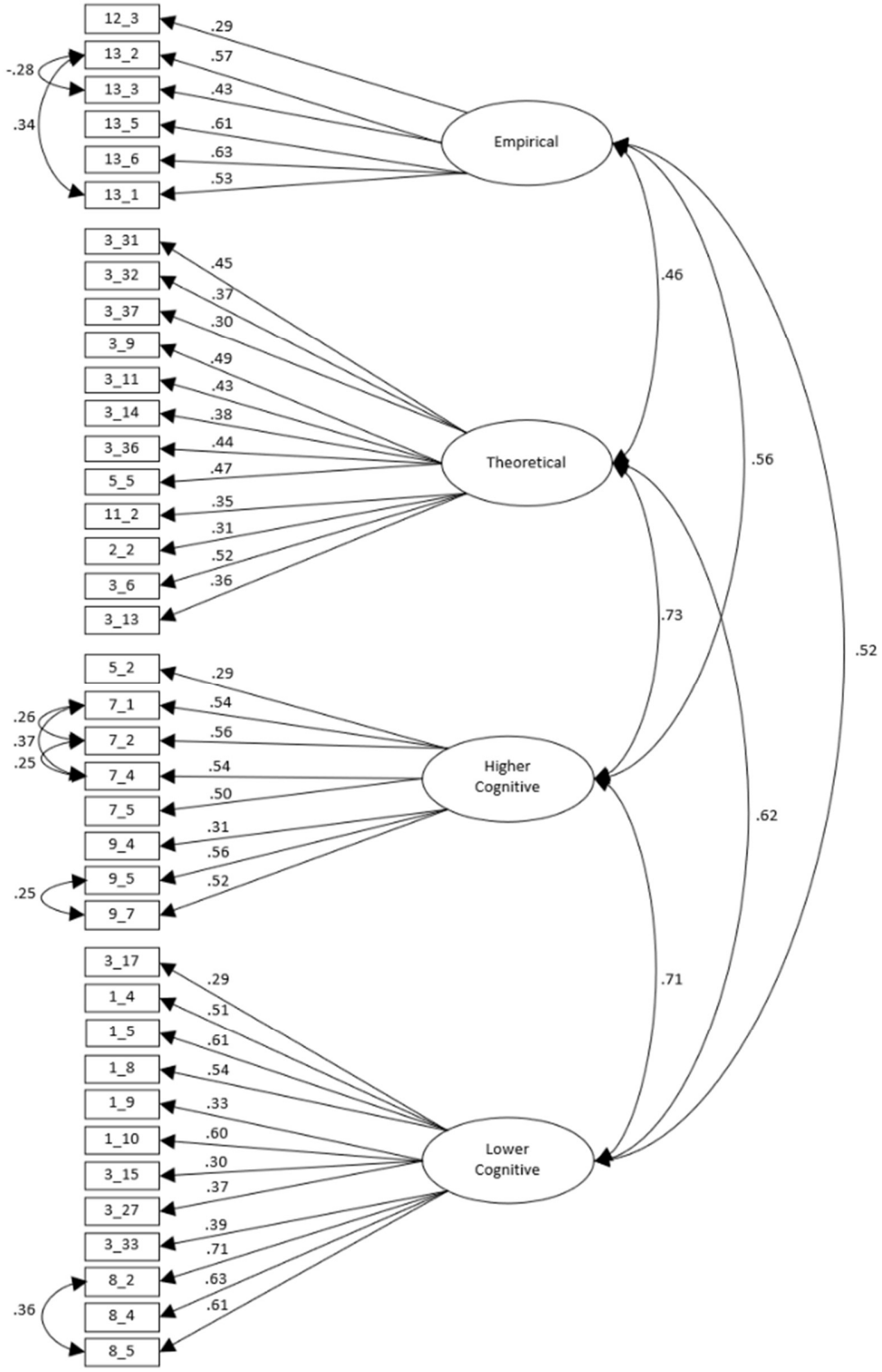
Next, we conducted a statistical analysis of the VAKAVA exam data to investigate if it produced a factor structure that would be comparable to the categories obtained with the qualitative content analysis and interpretable by the RBT (RQ2). EFA with the WLSMV estimator and oblique geomin rotation was conducted on all 106 items. We tested alternative factor solutions with one to five factors. Based on the fit indices, significance of factor loadings, and choosing items that clearly loaded on only one factor, the four-factor model with 62 items was selected. We used this model in further CFA analysis (see the Appendix for the factor loadings in the exploratory factor model).

Preliminary analysis indicated that at least some of the factors were interpretable according to the RBT. The fit of the model was acceptable ($\chi^2(5147)=16039$, $p < .001$; TLI=0.90; CFI=0.90; RMSEA=.02 (90% confidence interval [CI]: .018–.019); SRMR=.03).

In CFA, some items had poor factor loadings, and the model fit was unsatisfactory. The model was further modified by removing items with poor loadings (< .30; altogether, 24 items were excluded). Modification indices were also examined to determine whether the model fit improved when a particular path was added. Consequently, several residuals within the factors were allowed to correlate. In the final model, all factor loadings were significant, and each item loaded significantly on one factor. The four latent factors were allowed to correlate, with intercorrelations ranging from .46 to .73. The model fit was good ($\chi^2(652)=3005$, p <. 001; TLI=0.93; CFI=0.93; RMSEA=.02 (90% CI: .02–.03); SRMR=.05). Figure 3 shows the four factors with standardized factor loadings. These factors were named Higher Cognitive processing items, Lower Cognitive processing items, Theoretical items, and Empirical items.

Figure 3

Confirmatory Factor Model of the Multiple-choice Entrance Exam. All Factor Loadings and Correlations Are Significant at p < .001.



The factor for Higher Cognitive processing included eight items, which were all coded into the higher-order cognitive category in the qualitative content analysis (CR=.70). Additionally, the source material for all eight items was the theoretical paper. All of them concerned applying theoretical concepts to written examples of realistic situation descriptions in school life or society.

The factor for Lower Cognitive processing included 12 items that were coded into the lower-order cognitive skills category (CR=.80). The source material for these items was mostly the empirical paper. Alternatively, the items were based on a section of the theoretical paper that presented an example of an empirical study. The items were explicitly rephrased from the contents of the papers, and answering them required only finding the right short section of the paper and superficially comparing the item to the source material.

The factor for Theoretical items included 12 items (CR=.70). The source material for these items was the theoretical paper, the theoretical section of the empirical paper (e.g., the section presenting the paradigm), or the empirical paper in its entirety, thus requiring comprehension of the entire paper (e.g., comparing the findings of the empirical paper to the constructs in the theoretical paper).

The factor for Empirical items included six items (CR=.71). The source material for these items was the empirical paper. In these items, the applicants were asked to compare written descriptions of school life to the empirical findings of the research paper.

The descriptive statistics and correlations between the four factors are presented in Table 3.

Table 3

Means and Standard Deviations of the Sum Variables on the Items of the Factors and Correlations between the Latent Constructs

	Mean	*SD*	1	2	3	4
1 Empirical	0.37	0.46	-			
2 Theoretical	0.07	0.37	.46***	-		
3 Higher Cognitive	0.37	0.42	.56***	.73***	-	
4 Lower Cognitive	0.81	0.21	.52***	.62***	.71***	-
Exam total score	39.71	19.72	.65***	.90***	.95***	.90***

*** $p < .001$

The applicants obtained the highest scores in Lower Cognitive processing items, which were thus deemed relatively easy for the applicants. The Empirical and Higher Cognitive processing items were clearly more difficult for the applicants, while the mean of the most difficult Theoretical items approached zero (.07), indicating that very few applicants succeeded in these items. The repeated measures ANOVA indicated that the differences in the mean scores between the factors were statistically significant ($F(3, 6071)=8961.40$, $p < .001$). Theoretical and Higher and Lower Cognitive processing items were very strongly correlated with the total score in the VAKAVA exam, indicating that those applicants who mastered these items well were very likely to succeed in the entire exam as well.

The linear regression was significant ($\chi^2(4)=309.72$, $p < .001$) and indicated that the four factors were able to predict the variation in the total score of the written exam completely ($R^2=1.0$). The most significant predictor was the Theoretical factor (β=.59, *S.E.*=.02, $p < .001$). The Lower Cognitive factor predicted the VAKAVA score moderately (β=.40, *S.E.*=.02, $p < .001$). The predictions of the total score by the Empirical (β=.12, S.E.=.01, $p < .001$) and Higher Cognitive (β=.10, *S.E.*=.02, $p < .001$) factors were weaker but statistically significant.

7.3 Relations to the Matriculation Exam: Structural Model

We examined how the four factors were related to the ME scores (RQ3). We estimated all correlations between the four factors and the ME scores (Table 4). The fit of the model was good (χ^2(686)=2341–2415, p <. 001; TLI=0.92; CFI=0.91; RMSEA=.03 (90% CI: .02–.03); SRMR=.05).

Table 4

Correlations between the ME Grades and the Item Scores of the Latent Factors

	N	Empirical	Theoretical	Higher Cognitive	Lower Cognitive
ME average grade	3994	.33***	.45***	.43***	.37***
Mother tongue	3891	.24***	.35***	.29***	.26***
Mathematics adv	862	.15**	.18**	.14**	.08
Mathematics basic	2208	.17***	.29***	.23***	.18***
Psychology	1572	.18***	.24***	.24***	.25***
Health education	1791	.20***	.23***	.24***	.22***
Social studies	781	.05**	.16***	.09***	.12***
*Significance *** p < .001, ** p < .01*					

The correlations between the factors and the ME scores ranged from weak to moderate. The ME average grade correlated moderately with all the factors (.33–.45). Mother tongue and mathematics (advanced and basic) correlated most strongly with the Theoretical factor. Psychology and health education had the highest correlations with the Lower Cognitive, Higher Cognitive, and Theoretical factors. Advanced mathematics did not have a significant correlation with the scores of the Lower Cognitive factor, and the correlations with other factors were also weak. Basic mathematics correlated especially with the Theoretical and Higher Cognitive factors. The correlations of social studies with any factor were very weak.

Compared to the correlations, the linear regression analysis more clearly showed the differences between the factors in relation to the ME grades. The regression analyses (Table 5)

indicated that the Theoretical factors, as well as the Empirical and Higher Cognitive factors, predicted the ME average grade. The relations were weak to moderate but significant, and together, they predicted 24% of the variance in the ME average grade.

Table 5

Regression Analysis Summary for Entrance Exam Dimensions Predicting Matriculation Exam Grades (For the sample size for each variable, see Table 3.)

	Predictor								
	Empirical		Theoretical		Higher Cognitive		Lower Cognitive		
Variable	*β (S.E.)*	*p*	*β (S.E.)*	*p*	*β (S.E.)*	*p*	*β (S.E.)*	*p*	R^2
ME average grade	.12*** (.03)	<.001	.29*** (.04)	<.001	.12* (.05)	.03	.06 (.04)	.18	.24
Mother tng	.10*** (.03)	<.001	.28*** (.04)	<.001	.01 (.06)	.93	.04 (.04)	.39	.14
Math Adv	.12 (.07)	.08	.17* (.09)	.047	.03 (.13)	.82	-.10 (.10)	.29	.04
Math Basic	.06 (.04)	.14	.26*** (.05)	<.001	.03 (.07)	.73	-.02 (.06)	.75	.09
Psychology	.05 (.05)	.31	.13* (.06)	.04	.03 (.09)	.73	.13 (.07)	.06	.08
Health Ed	.09 *(.04)	.04	.11 (.06)	.06	.06 (.08)	.42	.07 (.06)	0.26	.07
SocialStudies	-.07 (.07)	.33	.21* (.09)	.02	-.01 (.11)	.91	.26** (.08)	.001	.14

*Significance *** p<.001, ** p<.01, * p<.05*

The Empirical and Theoretical factors predicted the mother tongue exam grade with statistical significance, explaining 14% of the variance. Mathematics (advanced and basic) and psychology grades were all predicted by the Theoretical factor only, whereas the health education grade was predicted by the Empirical factor. Social studies was the only school subject that was predicted by Lower Cognitive skills with statistical significance. Together with the Theoretical factor, they explained 14% of the variance in the ME grade in social studies.

8. **Discussion**

8.1 General Discussion

This research was an exploratory study examining the content and structure of the Finnish multiple-choice admission test, the VAKAVA exam, resulting in reflections on the pros and cons of this method of ITE student selection. The qualitative analysis of the cognitive processing skills required in the VAKAVA exam indicated that its multiple-choice items assessed both lower- and higher-order cognitive levels of the RBT. The share of items requiring lower-order cognitive processing with Remembering and Understanding was 60%, whereas that of items requiring higher-order processing with Analyzing and Applying accounted for 40%. These findings are in line with prior studies from other fields of higher education (Palmer & Devitt, 2007; Zheng et al., 2008). The exact desired distribution of low- or high-level tasks in an admission test is difficult to determine, but it is generally thought that higher education exams should preferably require higher-order thinking to reflect the requirements of the profession (Masters et al., 2001; Tarrant et al., 2006). The findings further indicated that applicants obtained, on average, significantly higher scores on lower-level than higher-level items. This expected finding indicates that the former items were easier for the applicants.

The statistical analyses of the VAKAVA exam data produced a factor structure that was interpretable by the RBT. Two factors, Higher Cognitive and Lower Cognitive items, included the same items that were previously identified in the qualitative analysis, although in the EFA and CFA, only part of the items loaded significantly on these two factors. To our knowledge, the present study is among the first to investigate the correspondence of qualitative and quantitative categorizations of multiple-choice items using the RBT as a conceptual framework. The findings

indicate that qualitative RBT categories, based on a systematic interpretation and classification of items to identify their cognitive levels, can also be found by statistically extracting underlying latent variables from a large number of observed variables. Although a relatively large number of items were excluded from the final factor model because of low factor loadings, our study supported the partial similarity between the conceptual and factor analytic categorizations of cognitive processing skills. Our study also corroborated prior evidence indicating that the RBT is a useful taxonomy for analyzing and categorizing the level of cognitive processing in testing and exams (Newton & Martin, 2013).

The exploratory and confirmatory factor modeling yielded two additional factors, which were named Empirical and Theoretical factors. This result suggests that the empirical items based on concrete observations and empirical data and, on the other hand, the theoretical items based on theories and concepts (Jaakkola, 2020) had systematic inter-dependence that could be explained by corresponding latent factors. The items in the Theoretical factor were, on average, the most difficult ones, which can be explained by their focus on scientific paradigms and the philosophy of science. As an explanation, ITE applicants may have been unfamiliar with reading theoretical scholarly papers, and their argumentative style and technical language may have caused difficulties in responding to the questions (van Lacum et al., 2012; Yarden, 2009), whereas it was easier for the applicants to extract the correct knowledge from an empirical article. No factors similar to Empirical and Theoretical have been identified in previous studies, presumably because studies have typically classified the cognitive skills required in multiple-choice tests with qualitative content analysis rather than with EFA and CFA, which allow the detection of hidden

patterns and their verification in the data. These findings underscore the importance of the mixed methods approach adopted in the current study.

Our findings further showed that the four factors correlated weakly or moderately with the ME scores. The highest correlations were at .43 and .45 and were found between Higher Cognitive and Theoretical factors and the ME average grade. These correlations were at the lower end of the range reported in previous literature (Metsäpelto et al., 2019, Räihä, 2010; Utriainen et al., 2012). The correlations of separate subjects with the four factors were in a somewhat lower range.

The strongest correlations were found for mother tongue, which correlated moderately particularly with Theoretical and Higher Cognitive factors. This finding is likely explained by the fact that mother tongue develops students' analytical and thinking skills along with textual skills, which support applicants in reading, analyzing, and applying knowledge in scientific articles. However, we suggest consideration of whether entrance exams favor native Finnish speakers, as teachers' cultural diversity should be supported and acknowledged already in admission processes (Klassen et al., 2020).

We also found that advanced mathematics had only weak correlations with the four factors. This was surprising because learning mathematics, particularly the advanced syllabus, requires strong skills in logical thinking and problem solving. Previous research indicates that the numerical skills of teacher students might be the best cognitive predictor of teacher effectiveness in the future (Bardach & Klassen, 2020). However, both research articles were qualitative and did not require the interpretation of statistical or numerical information, which may, in part, explain this finding.

The findings based on linear regression showed that the proportion of variance in the ME average grade explained by the four factors was .24. This finding indicates that a quarter of the variation in the ME average grade was explained by higher- and lower-order cognitive processing skills and the specific skills required to respond to items based on theoretical and empirical articles. The strongest predictor was applicants' score in the Theoretical factor, which reflects their skills to understand and compare theories. When predicting ME grades in singular subjects, the Theoretical factor was almost always the only significant predictor. This is understandable, as the current national curriculum for Finnish high schools underlines the multidisciplinary and research-based nature of psychology, health education, and social studies, as well as their goal of developing students' ability to gain a deep understanding of reflective information (Finnish National Agency for Education, 2019). Together, the findings linking the four factors extracted from the VAKAVA exam and ME grades indicated that the ME average grade had the greatest overlap with the skills needed to succeed in the VAKAVA exam and that the Theoretical factor stood out from the other factors with its stronger link to ME grades.

8.2 Practical Implications

The results of this study showed that the source materials played a significant role in the multiple-choice question format, as they largely determine the questions that can be formulated based on them. The qualitative and quantitative analyses indicated that the theoretical article allowed the design of particularly demanding questions that required higher-order cognitive processing and were difficult for the applicants, thus increasing the discriminative power of the exam and helping to identify those applicants with strong cognitive processing skills. On the other hand, the items based on the empirical article mainly required lower-order cognitive processing

skills. Therefore, the use of such an article as the sole source material could place too much emphasis on identifying trivial factual information, which has been a concern regarding the multiple-choice question format (Masters et al., 2001; Van der Vleuten, 1996).

Because of the large share of items assessing lower-order processing, it was possible to receive scores from the VAKAVA exam that were high enough to proceed to the next phase of student selection in some ITE programs without responding correctly to any items requiring higher-order cognitive skills. These findings underscore the need for the faculties creating the entrance exams to apply the knowledge to the content and structure of multiple-choice exams and to make a determined effort to increase the number of items requiring higher-order processing skills. Higher-order cognitive skills are essential in the teaching profession (Metsäpelto et al., 2021), as Finnish teacher students are supposed to absorb the epistemology and ontology of all elementary school subjects and be able not only to consume but also produce pedagogical knowledge (Krzywacki et al., 2015).

The selection of source materials also affects what other skills, in addition to lower- or higher-level processing skills, the applicant needs to respond to the exam. For instance, the questions in the VAKAVA exam are based on scholarly articles, thereby also requiring scientific reading skills. If the mathematical skills of the prospective teachers are considered critical for future teachers, the materials should preferably include quantitative content. Therefore, we recommend considering the broader set of desired skills required for the exam and selecting source materials based on this consideration. Of note, because good performance in the VAKAVA exam requires many cognitive skills, and the exam is based on the source material distributed on-site, it is difficult to prepare for the VAKAVA in advance. However, reading and analyzing diverse

scientific texts that contain academic argumentation and reasoning is likely to help applicants perform well in the exam.

8.3 Limitations and Methodological Reflections

The focus of this study, the VAKAVA exam, is used as an admission tool in student selection, so it is re-designed every year. Although the process of designing the exam is similar from year to year, the actual content and source materials vary annually. Therefore, the extent to which the findings (e.g., the distribution of items representing lower- and higher-order processing skills) of the study are reproducible and represent the exam over a longer period of time is unclear. Future studies should aim to replicate our findings.

The key focus of our study was to examine the cognitive level of items in the exam, which was accomplished by categorizing multiple-choice questions using the RBT as a conceptual framework. However, it should be noted that we did not have access to the actual mental processing of the applicants, so the specification of the required cognitive processing skills was necessarily an approximation of such skills. To avoid overestimating the proportion of items classified to a higher-order cognitive level, we classified the multiple-choice items including elements of higher-order cognitive processing at a lower level if it was sufficient to solve the problem. Therefore, we believe that our evaluation of the cognitive level of the VAKAVA exam was realistic.

We adopted a mixed methods approach, as there was no previous structural validation or background theory behind the VAKAVA exam. Therefore, the factor modeling was exploratory by nature, even when conducted with CFA. Achieving an excellent model fit was not possible, and somewhat lower fit indices were accepted (Hair et al., 2011). Triangulating qualitative and

quantitative methods was, however, essential to investigate the content and structure of the exam and allow theoretical interpretations and comparability to findings from prior studies (cf. Newton & Martin, 2013). Finally, our VAKAVA exam data did not include background information about the applicants. We do not know whether the findings are moderated by important background factors, such as the applicants' gender, age, or prior education. These issues remain to be investigated in future studies.

8.4 Conclusion

This study provides novel insights for understanding and improving ITE admission processes. The VAKAVA exam is an optional path to ITE for applicants whose ME grades are not high enough to proceed to the second phase of the admissions process. A very high correlation between VAKAVA scores and ME grades would indicate that those who failed in the ME would also fail in the VAKAVA exam. However, the moderate correlation observed in this study shows that some applicants succeed well in one and fail in the other, thus providing a real alternative path for proceeding in the admissions process. The weak or moderate correlations between VAKAVA exam scores and ME grades are likely explained by the different skill sets required for each test. While the ME assesses students' learning outcomes across several years in upper secondary school, requiring both cognitive and non-cognitive skills (e.g., persistence; Kupiainen et al., 2016), the VAKAVA exam assesses applicants' cognitive processing skills that have links to fast processing speed, memory, and reasoning abilities (Kail, 2000). In future, investigating the two-phase admission process as a whole with the non-cognitive phase included in the examination would be important to understand how current selection tools work, what kinds of

students they select for teacher education and, ultimately what is the predictive validity of the admission phase to student teachers' performance in ITE and in teacher profession.

The current study is based on a total population of 6074 ITE applicants, thus making this research unique and the findings generalizable. Specifically, it provides new findings about the multiple-choice format in student selection for ITE and has produced critical information for development work on admission processes (Baghaei et al., 2020). As such, it is connected to an increasing number of studies that have recently investigated different approaches to select students for ITE in different international contexts and educational systems (Klassen & Kim, 2021). These include situational judgment tests (Bardachet al., 2021) and multiple mini-interviews (Metsäpelto et al., 2022). The accumulating evidence base indicates that cognitive and non-cognitive attributes assessed at the student selection phase significantly predict student teachers' performance in ITE, such as in practicums (Klassen & Kim, 2019), signaling the importance of continuing efforts to investigate student selection for ITE in the future. Our study strengthens the research basis of student selection, but the results can also be applied more broadly when designing and applying multiple-choice questions in educational assessment.

References

Baghaei, S., Bagheri, M. S., & Yamini, M. (2020). Analysis of IELTS and TOEFL reading and listening tests in terms of Revised Bloom's Taxonomy. *Cogent Education, 7*(1). https://doi.org/10.1080/2331186X.2020.1720939

Bardach, L., & Klassen, R.M. (2020). Smart teachers, successful students? A systematic review of the literature on teachers' cognitive abilities and teacher effectiveness. *Educational Research Review, 30*, https://doi.org/10.31234/osf.io/nt7v9

Bardach, L., Klassen, R.M. & Perry, N.E. (2022). Teachers' Psychological Characteristics: Do They Matter for Teacher Effectiveness, Teachers' Well-being, Retention, and Interpersonal Relations? An Integrative Review. E*ducational Psychology Review, 34*, 259–300. https://doi.org/10.1007/s10648-021-09614-9

Bardach, L., Rushby, J.V., Kim, L.E., & Klassen, R.M. (2021). Using video-and text-based situational judgement tests for teacher selection: A quasi-experiment exploring the relations between test format, subgroup differences, and applicant reactions. *European Journal of Work and Organizational Psychology, 30*(2), 251–264. https://doi.org/10.1080/1359432X.2020.1736619

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Dordrecht: Springer.

Bloom, B.S. (1956). *Taxonomy of educational objectives. Vol. 1: Cognitive domain*. New York: McKay.

Case, S. M., & Swanson, D. B. (1998). *Constructing written test questions for the basic and clinical sciences* (2nd ed., pp. 22-25). Philadelphia: National Board of Medical Examiners.

Campbell, C.M., Michel, J.O., Patel, S. & Gelashvili, M. (2019). College Teaching from Multiple Angles: A Multi-trait Multi-method Analysis of College Courses. *Research in Higher Education, 60*, 711–735. https://doi.org/10.1007/s11162-018-9529-8

Cross, R., & O'Loughlin, K. (2013). Continuous assessment frameworks within university English Pathway Programs: realizing formative assessment within high-stakes contexts. *Studies in Higher Education, 38*(4), 584–594. https://doi-org.libproxy.helsinki.fi/10.1080/03075079.2011.588694

Darling-Hammond, L. (2017). Teacher education around the world: What can we learn from international practice? *European Journal of Teacher Education, 40*(3), 291–309, https://doi.org/10.1080/02619768.2017.1315399

DiDonato-Barnes, N., Fives, H., & Krause, E.S. (2014). Using a Table of Specifications to improve teacher-constructed traditional tests: an experimental design. *Assessment in Education: Principles, Policy & Practice, 21*(1), 90–108 https://doi.org/10.1080/0969594X.2013.808173

Dwyer, C., Hogan, M., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity, 12*, 43–52. https://doi.org/10.1016/j.tsc.2013.12.004

Finnish National Agency for Education (2019). *National Core Curriculum for General Upper Secondary Education.*

https://www.oph.fi/sites/default/files/documents/lukion_opetussuunnitelman_perusteet_2019.pdf

Fuller, U., Johnson, C., Ahoniemi, T., Cukierman, D., Hernán-Losada, I., Jacková, J., Lahtinen, E., Lewis, T., Thompson, D., Riedesel, C. & Thompson, E. (2007). Developing a computer science-specific learning taxonomy. *ACM SIGCSE Bulletin, 39*, 152-170. https://doi.org/10.1145/1345375.1345438.

Hair, J.F., Ringle, C.M., & Sarstedt, M. (2011) PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice, 19*(2), 139-152. https://doi.org/10.2753/MTP1069-6679190202

Hanna, W. (2007). The new Bloom's taxonomy: Implications for music education. *Arts Education Policy Review, 108*(4), 7-16. https://doi.org/10.3200/AEPR.108.4.7-16

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Routledge.

Hammerness, K., Ahtiainen, R., & Sahlberg, P. (2017). *Empowered educators in Finland: how high-performing systems shape teaching quality*. San Francisco: Jossey-Bass.

Hoe, S. (2008). Issues and Procedures in Adapting Structural Equation Modeling Technique. Quantitative Methods Inquires. *Journal of Applied Quantitative Methods, 3*(1), 76–83.https://ink.library.smu.edu.sg/sis_research/5168

Hu, L. & Bentler, P. (1995). Evaluating model fit. In: R. Hoyle (Eds.), *Structural equation modeling: Concepts, issues, and applications* (p. 76–99). Thousand Oaks: Sage.

Ingvarson, L. (2013). Recruitment and selection in teacher education. In L. Ingvarson, J. Schwille, M.T. Tatto, G. Rowley, R. Peck, & S.L. Senk (2013). *An analysis of teacher education context, structure, and quality-assurance arrangements in TEDS-M countries: Findings from the IEA teacher education and development study in mathematics (TEDS-M)* (pp. 165–209). https://files.eric.ed.gov/fulltext/ED545244.pdf#page=166

Jaakkola, E. (2020). Designing conceptual articles: four approaches. *AMS review, 10*(1), 18–26. https://doi.org/10.1007/s13162-020-00161-0

Jansen, T., & Möller, J., (2022). Teacher judgments in school exams: Influences of students' lower-order-thinking skills on the assessment of students' higher-order-thinking skills. *Teaching and Teacher Education, 111*. https://doi.org/10.1016/j.tate.2021.103616.

Jensen J.L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the Test…or Testing to Teach: Exams Requiring Higher Order Thinking Skills Encourage Greater Conceptual Understanding. *Educational Psychology Review, 26*(2), 307–329. https://doi.org/10.1007/s10648-013-9248-9

Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction. In R. Ruddell & N. Unrau (Eds.), *Theoretical models and processes of reading*, (5th ed, pp. 1270-1328). Newark: International Reading Association.

Kail, R. (2000). Speed of information processing: Developmental change and links to intelligence. *Journal of School Psychology*, 38(1), 51-61.

Klassen, R.M. & Kim, L.E. (2019). Selecting teachers and prospective teachers: A meta-analysis. *Educational Research Review, 26*, 32-51. https://doi.org/10.1016/j.edurev.2018.12.003

Klassen, R.M., Kim, L.E., Rushby, J.V., & Bardach, L. (2020). Can we improve how we screen applicants for initial teacher education? *Teaching and Teacher Education, 87*. https://doi.org/10.1016/j.tate.2019.102949.

Krathwohl, D. R., & Anderson, L. W. (2001*). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Krzywacki, H., Lavonen, J., & Juuti, K. (2015). There are No Effective Teachers in Finland— Only Effective Systems and Professional Teachers. In O-S. Tan, & W-C. Liu (Eds.), *Teacher Effectiveness: Capacity Building in a Complex Learning Era* (pp. 79-103). Centage learning.

Kuhn, M. & Stahl, S. (2003), Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, 95(1), 3-21. http://dx.doi.org/10.1037/0022-0663.95.1.3.

Kuncel, N.R., Hezlett, S.A. & Ones, D.S. (2001). A Comprehensive Meta-Analysis of the Predictive Validity of the Graduate Record Examinations. *Psychological Bulletin, 127*(1), 162-181. https://doi.org/10.1037/0033-2909.127.1.162

Kupiainen, S., Marjanen, J. M., & Hautamäki, J. J. (2016). The problem posed by exam choice on the comparability of results in the Finnish matriculation examination. *Journal for Educational Research, 8*(2), 87–106.

Lennox, R., Hepburn, K., Leaman, E., & van Houten, N. (2020). 'I'm probably just gonna skim': an assessment of undergraduate students' primary scientific literature reading approaches. *International Journal of Science Education, 42*(9), 1409-1429. https://doi.org/10.1080/09500693.2020.1765044

Malinen, O.-P., Väisänen, P., & Savolainen, H. (2012). Teacher education in Finland: a review of a national effort for preparing teachers for the future. *Curriculum Journal, 23*(4), 567–584. https://doi.org/10.1080/09585176.2012.731011

Mankki, V., Mäkinen, M., & Räihä, P. (2020). Teacher educators' predictability and student selection paradigms in entrance examinations for the Finnish Primary School Teacher Education programme. *European Journal of Teacher Education, 43*(2), 151–164. https://doi.org/10.1080/02619768.2019.1672653

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education, 40*(1), 25-32. https://doi.org/10.3928/0148-4834-20010101-07

Matriculation Examination Board (2022, June 9) *Matriculation Examination*. https://www.ylioppilastutkinto.fi/en/matriculation-examination

Metsäpelto, R-L., Viljaranta, J., Tuominen, H., Aunola, K., Poikkeus, A-M., & Mullola, S. (2019). Ylioppilastutkinnon, tavoiteorientaatioiden ja muiden motivaatiotekijöiden yhteys luokanopettajakoulutukseen hakeneiden menestymiseen VAKAVA-valintakokeessa. *Kasvatus - The Finnish Journal of Education, 50*(2), 136–148.

Metsäpelto, R.-L., Utriainen, J., Poikkeus, A.-M., Muotka, J., Tolvanen, A., & Warinowski, A. (2022). Multiple mini-interviews as a selection tool for initial teacher education admissions. *Teaching and Teacher Education, 113*. https://doi.org/10.1016/j.tate.2022.103660

Metsäpelto, R-L., Poikkeus, A-M., Heikkilä, M., Husu, J., Laine, A., Lappalainen, K., Lähteenmäki, M., Mikkilä-Erdmann, M., & Warinowski, A. (2021). A Multidimensional Adapted Process Model of Teaching. *Educational Assessment, Evaluation and Accountability*. https://doi.org/10.1007/s11092-021-09373-9

Moè, A., & Katz, I. (2020). Self-compassionate teachers are more autonomy supportive and structuring whereas self-derogating teachers are more controlling and chaotic: The mediating role of need satisfaction and burnout. *Teaching and Teacher Education, 96.* https://doi.org/10.1016/j.tate.2020.103173

Muthén, L.K. & Muthén, B.O. (1998-2017). *Mplus User's Guide. Eighth Edition*. Los Angeles: Muthén & Muthén.

Neiro, J., & Johansson, N. (2020). The Finnish matriculation examination in biology from 1921 to 1969 – trends in knowledge content and educational form. *LUMAT: International Journal on Math, Science and Technology Education, 8*(1), 162–199. https://doi.org/10.31129/LUMAT.8.1.1376

Newton, G., & Martin, E. (2013). Blooming, SOLO Taxonomy, and Phenomenography as Assessment Strategies in Undergraduate Science Education. *Journal of College Science Teaching, 43*(2), 78–90. http://www.jstor.org/stable/43631075

Niessen, A.S., Rob, R.M., & Tendeiro, J.N. (2018). Admission testing for higher education: A multi-cohort study on the validity of high-fidelity curriculum-sampling tests. *PLoS One, 13*(6) http://dx.doi.org/10.1371/journal.pone.0198746

Niessen, A.S., Meijer, R.R., Tendeiro, J.N. (2017). Applying organizational justice theory to admission into higher education: Admission from a student perspective. *International Journal of Selection and Assessment, 25*(1), 72-84. https://doi.org/10.1111/ijsa.12161

OECD (2021). *The Assessment Frameworks for Cycle 2 of the Programme for the International Assessment of Adult Competencies*. OECD Skills Studies. OECD Publishing: Paris. *https://doi.org/10.1787/4bc2342d-en*.

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Medical Education, 7*(1), 1-7. https://doi.org/10.1186/1472-6920-7-49

Parvaneh, S. (2020). A model of interplay between student English achievement and the joint affective factors in a high-stakes test change context: Model construction and validity. *Educational*

Assessment, Evaluation and Accountability, 32(3), 335-371. http://dx.doi.org/10.1007/s11092-020-09326-8

Peltola, M., Suorsa, T., Karhu, J. & Soini, H. (2020). Huoli kytkeytyy osallisuuden rajapintoihin: Ammattilaisen arki oppilashuollon näyttämöillä. [Concern is linked to the interfaces of inclusion: The everyday life of a professional in the student welfare scene.] *Aikuiskasvatus, 40*(2), 127–138.

Räihä, P. (2010). Vakava-hankkeesta ei tullutkaan uuden ylioppilaan pelastajaa. *Kasvatus, 41*(3), 213-225.

Raatikainen, P. (2015). Ymmärtäminen ja selittäminen ihmistieteissä. [Understanding and explaining in social sciences.] *Kasvatus, 46*(3), 281–286.

Richardson, M., Abraham, C. & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353-387. https://doi.org/10.1037/a0026838

Rouet, J., Britt, M., & Potocki, A. (2019). "Multiple-text comprehension". In J. Dunlosky & K. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education (Cambridge Handbooks in Psychology, pp. 356-380)*. Cambridge University Press, Cambridge. http://dx.doi.org/10.1017/9781108235631.015.

Stone, C. A., & Yeh, C. C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. *Educational and Psychological Measurement, 66*(2), 193-214. https://doi.org/10.1177/2158244016674513

Stringer, J.K., Santen, S.A., Lee, E., Rawls, M., Bailey, J., Richards, A., Perera, R.A. & Biskobing, D. (2021). Examining Bloom's Taxonomy in Multiple Choice Questions: Students' Approach to Questions. *Medical Science Educator, 31*, 1311–1317. https://doi.org/10.1007/s40670-021-01305-y

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today, 26*(8), 662-671. https://doi.org/10.1016/j.nedt.2006.07.006

Taxer, J. L., & Gross, J. J. (2018). Emotion regulation in teachers: The "why" and "how". *Teaching and Teacher Education, 74,* 180-189. https://doi.org/10.1016/j.tate.2018.05.008

Thomson, D., Cummings, E., Ferguson, A. K., Moizumi, E. M., Sher, Y., Wang, X., Broad, K., & Childs, R. A. (2011). A role for research in initial teacher education admissions: A case study from one Canadian university. *Canadian Journal of Educational Administration and Policy, 121*, 1–23. https://cjcrcc.ucalgary.ca/index.php/cjeap/article/view/42818

Thompson, A.R. & O'Loughlin, V.D. (2015). The Blooming Anatomy Tool (BAT): A discipline-specific rubric for utilizing Bloom's taxonomy in the design and evaluation of assessments in the anatomical sciences. *Anatomical Sciences Education, 8*(6), 493-501. https://doi.org/10.1002/ase.1507

Thompson, A.R., Braun, M.W., & O'Loughlin, V.D. (2013). A comparison of student performance on discipline-specific versus integrated exams in a medical school course. *Advances in Physiology Education 37*(4), 370-376. https://doi.org/10.1152/advan.00015.2013

University of Jyväskylä (2020). Opetussuunnitelmat 2020-2023: Luokanopettajan kandidaatti- ja maisterikoulutus [Curriculums 2020-2023: The bachelor's and master's education of classroom teacher]. https://www.jyu.fi/edupsy/fi/laitokset/okl/opiskelu/luokanopettajakoulutus/opetussuunnitelmat-ja-opetusohjalmat

Utriainen, J., Kallio, E., & Tynjälä, P. (2012). Opiskelijavalintojen kehittäminen kasvatustieteessä: tutkimus- ja kehityshankkeen loppuraportti. *Työpapereita/Koulutuksen tutkimuslaitos, 28*.

Utriainen, J., Marttunen, M., Kallio, E., & Tynjälä, P. (2017). University Applicants' Critical Thinking Skills: The Case of the Finnish Educational Sciences. *Scandinavian Journal of Educational Research, 61*(6), 629–649. https://doi.org/10.1080/00313831.2016.1173092

van Lacum, E., Ossevoort, M., Buikema, H., & Goedhart, M. (2012). First experiences with reading primary literature by undergraduate life science students. *International Journal of Science Education, 34*(12), 1795–1821. https://doi.org/10.1080/09500693.2011.582654

Vanthournout, G., Donche, V., Gijbels, D. & Van Petegem, P. (2014). (Dis)similarities in research on learning approaches and learning patterns. In D. Gijbels, V. Donche, J.T.E. Richardson & J.D. Vermunt (Eds.), *Learning patterns in higher education. Dimensions and research perspectives*. London, UK: Routledge, 11– 32.

Vermunt, J.D. (2005). Relations between student learning patterns and personal and contextual factors and academic performance. *Higher Education, 49*, 205–234. https://doi.org/10.1007/s10734-004-6664-2

Vermunt, J.D., & Donche, V. 2017. A learning patterns perspective on student learning in higher education: State of art and moving forward. Educational Psychology Review 29, 269–299. https://doi.org/10.1007/s10648-017-9414-6

Van Der Vleuten, C.P.M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education, 1*(1), 41–67. https://doi.org/10.1007/BF00596229

Walczyk, J.J., Marsiglia, C.S., Johns, A.K., Bryan, K.S. (2004). Children's compensations for poorly automated reading skills. *Discourse Processes, 37*(1), 47-66, http://dx.doi.org/10.1207/s15326950dp3701_3.

Yarden, A. (2009). Reading scientific texts: Adapting primary literature for promoting scientific literacy. *Research in Science Education*, *39*(3), 307–311. https://doi.org/10.1007/s11165-009-9124-2

Zaidi, N., Grob, K., Monrad, S., Kurtz, J., Tai, A., Ahmed, A., Gruppen, L., & Santen, S. (2018). Pushing critical thinking skills with multiple-choice questions: Does Bloom's Taxonomy work? *Academic Medicine, 93*(6), 856-859. https://doi.org/10.1097/ACM.0000000000002087.

Zaidi, B., Grob, K.L., Yang, J., Santen, S.A., Monrad, S.U., Miller, J.M., & Purkiss, J.A. (2016). Theory, Process, and Validation Evidence for a Staff-Driven Medical Education Exam Quality Improvement Process. *Medical Science Educator, 26*(3), 331–336. https://doi.org/10.1007/s40670-016-0275-2

Zheng A.Y., Lawhorn J.K., Lumley T., & Freeman S. (2008). Application of Bloom's taxonomy debunks the "MCAT myth". *Science, 319*, 414–415.

Appendices

Appendix 1

The Factor Loadings of the EFA

	Factor					Factor			
Item	1	2	3	4	Item	1	2	3	4
1_6	**.45***	-.04	.14*	-.00	8_5	-.00	**.48***	.21*	.05*
2_1	**.48***	.02	.13*	.03	9_1	.03	**.38***	-.09*	-.02
2_2	**.33***	.05*	-.04*	.00	9_2	.09*	**.27***	-.10*	-.05*
2_5	**.34***	-.04	.00	-.03	9_4	.05*	**.36***	-.10*	-.01
3_6	**.36***	.11*	.23*	-.01	9_5	.02	**.60***	.04	-.03
3_9	**.42***	.06*	-.00	.10*	9_7	.02	**.54***	.02	.01
3_11	**.37***	.03	.00	.14*	1_4	.24*	.01	**.34***	.10*
3_12	**.22***	.03	.13*	-.02	1_5	.10*	.20*	**.37***	.14*
3_13	**.44***	.06*	-.15*	.02	1_8	.07*	-.03	**.55***	.17*
3_14	**.40***	-.05*	.17*	-.03	1_9	-.07*	-.09*	**.51***	.15*
3_25	**.38***	.05*	.07*	.03	1_10	.21*	.09*	**.33***	.23*
3_30	**.44***	-.08*	-.06*	.07*	2_6	.08*	.05	**.33***	-.08*
3_31	**.33***	.07*	.14*	.02	3_2	.01	.07*	**.32***	-.10*
3_32	**.42***	.11*	-.12*	-.02	3_10	.09*	.02	**.23***	.06*
3_35	**.50***	-.10*	-.21*	.07*	3_15	.06*	.04	**.38***	-.05*
3_36	**.34***	.09*	.16*	-.03	3_17	.02	.04	**.39***	.00
3_37	**.46***	-.11*	-.02	.03	3_22	-.43*	-.09*	**.56***	-.03
4_3	**.21***	-.03	-.04	-.02	3_27	-.20*	-.04	**.74***	.06*
5_5	**.28***	.16*	.08*	.06*	3_29	-.13*	.09*	**.25***	.06*
6	**.26***	-.00	.04	.04*	3_33	.06*	.06*	**.47***	-.03
11_2	**.29***	.08*	.02	.03	3_34	.02	.24*	**.41***	-.02
5_2	.11*	**.17***	.12*	.01	11_1	-.09*	.21*	**.29***	.24*
5_6	-.04*	**.10***	-.01	.05*	11_5	.13*	-.01	.05*	**.16***
7_1	.22*	**.47***	-.07*	.02	12_2	.05*	-.14*	.00	**.57***
7_2	-.03	**.55***	.13*	.01	12_3	.03	.02	.06*	**.36***
7_4	.16*	**.49***	-.02	.00	12_5	.02	.04*	-.04*	**.34***
7_5	.09*	**.37***	.10*	.04	13_1	.07*	.26*	-.16*	**.37***
7_7	-.09*	**.49***	.24*	-.01	13_2	.04*	.23*	-.03	**.39***
8_1	.04*	**.47***	.16*	.02	13_3	-.03	.14*	.06*	**.23***
8_2	-.09*	**.59***	.32*	.05	13_5	.00	.21*	.01	**.39***
8_4	.17*	**.36***	.11*	.09*	13_6	.05*	.26*	.01	**.39***