OmniaScience

Journal of Technology and Science Education
JOTSE, 2020 – 10(2): 259-270 – Online ISSN: 2013-6374 – Print ISSN: 2014-5349
https://doi.org/10.3926/jotse.1017

# ASSESSING FIRST-YEAR ENGINEERING STUDENTS' PRE-UNIVERSITY MATHEMATICS KNOWLEDGE: PRELIMINARY VALIDITY RESULTS BASED ON AN ITEM RESPONSE THEORY MODEL

**Yusuf F. Zakariya, H.K. Nilsen, Simon Goodchild, Kirsten Bjørkestøl**

University of Agder, Department of Mathematical Sciences (Norway)

*yusuf.zakariya@uia.no, hans.k.nilsen@uia.no, simon.goodchild@uia.no, kirsten.bjorkestol@uia.no*

*Received May 2020*
*Accepted June 2020*

**Abstract**

The importance of students' prior knowledge to their current learning outcomes cannot be overemphasised. Students with adequate prior knowledge are better prepared for the current learning materials than those without the knowledge. However, assessment of engineering students' prior mathematics knowledge has been beset with a lack of uniformity in measuring instruments and inadequate validity studies. This study attempts to provide evidence of validity and reliability of a Norwegian national test of prior mathematics knowledge using an explanatory sequential mixed-methods approach. This approach involves use of an item response theory model followed by cognitive interviews of some students among 201 first-year engineering students that constitute the sample of the study. The findings confirm an acceptable construct validity for the test with reliable items and a high-reliability coefficient of .92 on the whole test. Mixed results are found on discrimination and difficulty indices of questions on the test with some questions having unacceptable discriminations and require improvement, some are easy, and some appear too tricky questions for students. Results from the cognitive interviews reveal the likely reasons for students' difficulty on some questions to be lack of proper understanding of the questions, text misreading, improper grasping of word-problem tasks, and unavailability of calculators. The findings underscore the significance of validity and reliability checks of test instruments and their effect on scoring and computing aggregate scores. The methodological approaches to validity and reliability checks in the present study can be applied to other national contexts.

***Keywords –*** Prior knowledge, Item response theory, Mixed methods, Validity, Reliability.

***To cite this article:***

Zakariya, Y.F., Nilsen, H.K., Goodchild, S., & Bjørkestøl, K. (2020). Assessing first-year engineering students' pre-university mathematics knowledge: Preliminary validity results based on an item response theory model. *Journal of Technology and Science Education,* 10(2), 259-270.
https://doi.org/10.3926/jotse.1017

---

## 1. Introduction

Students' knowledge before a teaching-learning activity has been reported in diverse fields of studies to exert enormous influence in facilitating proper understanding of current learning materials. Many psychological theories (e.g., self-efficacy theory) have acknowledged and emphasised this strong predictive role of prior knowledge on the current learning outcomes (Bandura, 1997; Marton & Booth, 1997). The

correlation between prior academic achievement and students' performance in the presented tasks has been extensively reported in the literature. In a study of 60 first-year undergraduate students of accounting and economics reported by Duff (2004), prior academic achievement is found to correlate (r=.53) with academic performance positively, and it is the best among other predictors such as age and gender. This report is corroborated in a much larger sample longitudinal study in which prior academic achievement is also found to be the best, among other factors, in predicting 1,628 secondary school students' performance in reading and mathematics (Engerman & Bailey, 2006).

Furthermore, Ayán and García (2008) compare the efficacy of linear and logistic regression models in predicting 639 undergraduate students' performance, and both models favour prior academic achievement over other factors such gender and school location. In the same year, Hailikari, Nevgi and Komulainen (2008) conducted a special problem-solving mathematical assessment to determine students' prior knowledge in their study and its predictive power of academic performance. Their report shows that prior knowledge, coupled with previous academic success explained 55% of the variability observed in the performance of students on mathematics tasks. Similar results are also reported, elsewhere, (e.g., Casillas, Robbins, Allen, Kuo, Hanson & Schmeiser, 2012; Newman-Ford, Lloyd & Thomas, 2009; Richardson & Abraham, 2012).

Recently, a group of researchers Martin, Wilson, Liem and Ginns (2016) recorded mixed results on the prior knowledge predictive power of performance in a 2-year longitudinal study among university students. High school results as proxies for measuring prior knowledge correlate well with the performance at the beginning of their study while the ongoing semester course grades take the lead later. Though, this finding seems not contradictory to the earlier reported ones as both high school grades and the ongoing semester course grades still refer to prior academic achievement of the students in some sense. The findings reported by Aluko, Daniel, Oshodi, Aigbavboa and Abisuga (2018); Opstad, Bonesrønning and Fallan (2017) corroborate this point. Aluko et al. (2018) utilised more sophisticated statistical tools such as logistic regression and support vector machine learning to establish high correction between prior academic achievement and performance.

Despite the importance of prior knowledge and its correlation with students' performance, studies on psychometric properties of measures of engineering students' prior mathematics knowledge are scarce in the literature. As such, the primary purpose of the present study is to validate a prior knowledge of mathematics test (PKMT), owned by the Norwegian Mathematical Council, using an item response theory (IRT) model coupled with some cognitive interviews to extract detail information on likely reasons why some questions are challenging for students. The present study will not only provide empirical evidence for the validity of the PKMT but also offer pieces of advice to the Norwegian Mathematical Council towards an improvement of specific items on the test. Further, validation of the PKMT is also crucial for our ongoing relatively large-scale quantitative study on the contributions of prior mathematics knowledge, approaches to learning and self-efficacy on year-one engineering students' performance in mathematics at a Norwegian university. It is important to remark that the report presented in this article is preliminary and as such more studies are still ongoing in relating the scores of students on the prior knowledge of mathematics test to students' grades and other constructs.

The remaining part of the present article is arranged such that a conceptual framework is elucidated in the next section. The section was followed by another section where issues related to methodology, e.g., an overview of some specifics of the PKMT, sample of the study and procedure of data collection and analysis are presented. This is followed by a section where we present and discuss ensuing results from both the quantitative and the qualitative approaches to data analysis. The last section, before the reference list sheds more lights on the significant findings of the study, gives some concluding remarks and acknowledges the strengths and potential weaknesses of the study.

## 2. Conceptual Framework

### 2.1. The Conceptualisation of Prior Knowledge

There seems to be no agreement among educationists and psychologists on a definition of prior knowledge. Though, it is used to be captured as cognitive entry behaviour enshrined in Bloom's taxonomy (Bloom, 1976). The definition of Bloom's cognitive entry behaviour as "those prerequisite types of knowledge, skills, and competencies which are essential to the learning of a particular new task or set of tasks" (Bloom, 1976: page 122) has been criticised and considered outdated in some quarters (e.g. Dochy, De Rijdt & Dyck, 2002). In their review, Dochy et al. (2002) explicate many synonymous terms used to describe prior knowledge in the literature and consider their general interpretations to be "definitional snippets or vague statements" (Dochy et al., 2002: page 267). Thus, Dochy et al. (2002) propose and describe prior knowledge as:

> *The whole of a person's knowledge, which is as such dynamic in nature, is available before a certain learning task, is structured, can exist in multiple states (i.e. declarative, procedural and conditional knowledge), is both explicit and tacit in nature and contains conceptual and metacognitive knowledge components* (Dochy et al., 2002: page 267).

Another approach through which prior knowledge has been conceptualised is from an angle of domain-specific tasks or accomplishments. In this view, prior knowledge is seen as the level of knowledge related to a specific field being studied which varies distinctively depending on the relevance and the quality of the material currently under study (Dochy, 1996; Hailikari et al., 2008). Thus, prior knowledge in the present study refers to prior mathematics performance of students before they start their university education.

The notion of domain-specific prior knowledge seems to provide a basis for different indicators used in the literature to assess prior knowledge of the learners. There has been little coherence between various indicators used by educationists as proxies to quantify students' prior knowledge. This lack of uniformity can be linked to the type of studies, e.g. longitudinal (Engerman & Bailey, 2006), meta-analysis (Richardson & Abraham, 2012); students under study, e.g. university (Ayán & García, 2008), high school students (Casillas et al., 2012); the field of study, e.g. accounting (Duff 2004), mathematics (Hailikari et al., 2008), economics (Opstad et al., 2017), and architecture (Aluko et al., 2018). In several of these studies, researchers have used students' test scores on standardised tests, high school grades and entrance exams (Aluko et al., 2018; Casillas et al., 2012; Duff 2004; Newman-Ford et al., 2009) while others have used students previous semester/year grades (Ayán & García, 2008; Engerman & Bailey, 2006; Martin et al., 2016; Zakariya, 2016) or a special exam on problem-solving (Hailikari et al., 2008) to assess prior knowledge.

### 2.2. Study Setting

Students that are admitted into science and engineering courses at Norwegian universities have the freedom to choose between three routes and two endpoints for their mathematics studies at upper secondary schools (grades 11-13). The routes are practical mathematics (P-Mat) aiming at applications of mathematics, social science mathematics (S-Mat) and advanced mathematics for science and technology (R-Mat) and they can conclude their study of mathematics after two or three years at their upper secondary schools. The Norwegian Mathematical Council has consistently administered a prior knowledge of mathematics test (PKMT) to year-one university and college students since 1984. The PKMT aims to provide empirical evidence for monitoring of the basic knowledge of mathematics with a focus on undergraduate students following mathematics intensive programmes (e.g. engineering programmes) across universities and colleges in Norway. The PKMT is conducted every two years since 2001, and the latest was conducted in Autumn 2019. Prior to the year 2001, the test was conducted in 1984, 1986, 1999 and 2000. Accordingly, based on the results of the PKMT pieces of advice are offered by the Norwegian Mathematical Council to government agencies, Norwegian Research Council, universities, colleges, and other mathematics education stakeholders in Norway. However, it is apparent that some mathematics

educators and researchers in Norway have reservations about the validity of the PKMT. It is the opinion of the authors that some of these reservations could be traced to a lack of validation studies on the instrument which has motivated the present study.

## 3. Methodology

### 3.1. Measure

The PKMT has two main parts. The first part contains background information about the students such as gender, age, some information about the highest mathematics content followed in upper secondary schools, and some items on attitudes towards mathematics. The second part is a 16-item test on basic mathematics tasks that are developed based on secondary school (grades 8-10) curriculum. Items 1 and 2 have three parts each, items 9 and 11 have two parts each, while other items have only one part each to make a total of 22 questions on the test. Questions 9a, 11a, 11b, 14 and 15 are standard multiple-choice questions while others are short open-ended questions. Before the commencement of the present study, the PKMT is administered using paper and pencil format. Thus, we independently digitalised the test and administered it online under classroom supervision. Coincidentally, the Norwegian Mathematical Council also shifted to digital PKMT in the 2019 administration of the test at the national level. The use of calculators is not allowed, and it takes 40 minutes to complete the test, including the time to complete background information. Sample questions of the PKMT are not included in the present article for confidentiality reasons. However, questions on the test can be categorised into five clusters: (a) basic operations of addition, multiplication, division and ordering of fractions and decimals; (b) simple percentages, ratio, proportion and average speed; (c) solving linear equations and inequalities including an application of Pythagoras theorem; (d) reading a Cartesian graph, slope of a straight line, similar triangles and volume of solid shapes; and (e) word problems on writing, interpreting and solving linear/simultaneous equations. Further, some of the questions are discussed in parts in a way that they are not identifiable during the presentation of some interview transcripts.

### 3.2. Participants

A total of 201 year-one engineering students in a Norwegian university including 34 females and 167 males took the PKMT in Autumn 2019. The average age of the students is 20.64 years, with a minimum of 17 years and a maximum of 36 years. Appropriate consents are sought from the Norwegian Centre for Research Data (NSD) as well as individual students who took parts in the test. The students are made to understand that taking part in the study is entirely voluntary and that their refusal to give consent will not in any way affect their grades. They are promised that utmost confidentiality will be ensured in dealing with their data and that no student is identifiable during and after the study. The data used for the present study are completely anonymous and are available upon request from the corresponding author.

### 3.3. Data Analysis

The collected data are initially scored dichotomously using 1 point for a correct answer and 0 point for a wrong. The scored data are analysed using a quantitative method. A two-parameter IRT model was used to investigate item parametrisations such as item discriminating and difficulty indices as well as item reliability of the test. An IRT model is a framework that characterises a relation between examinee's ability or latent trait as measured by a scale and the examinee's responses to each item on the scale (DeMars, 2010). IRT models can be one-parameter, two-parameter, three-parameter, unidimensional (i.e., items measure a common latent trait) and multidimensional (i.e., items measure separate clusters of a latent trait) depending on the complexity of the scale. The basic notion of the two-parameter IRT model is that a subject's probability of getting an item correct is a monotonic increasing function (e.g., an exponential function) of two sets of parameters: (a) the location (item difficulty) on the latent trait (in our case, prior mathematics knowledge) to be measured; and (b) the slope (item discrimination) of item response function (IRF) otherwise known as item characteristic curve (ICC). Equation 1 presents a mathematical representation of a two-parameter IRT model.

$$P(X_i = 1|\theta, a_i, b_i) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \quad (1)$$

Equation 1 shows the probability (P) that a student with latent variable $\theta$ (competence on the PKMT) answers an item ($X_i = 1$) correctly which has both item difficulty and discrimination indices of $a_i$ and $b_i$ respectively and $e$ is an exponential function. The test scores on the PKMT is put on a metric determined by IRT model such that the group latent variable is normally distributed (mean = 0 and standard deviation = 1) with values ranging from -3.5 to +3.5. Each item discrimination index ($a_i$ , $i$ = 1, 2, …, 22) has the same metric as the latent variable ($\theta$) with values ranging over the set of real numbers. It measures the extent to which an item discriminates between students of low and high ability on the PKMT. Items with negative or less than 0.20 $a_i$'s have been recommended to be rejected while items with $0.20 \leq a_i\text{'s} < 0.40$ demonstrate appropriate discrimination and may be improved and items with $a_i\text{'s} \geq 0.40$ demonstrate good discrimination (DeMars, 2010; Ebel & Frisbie, 1991). However, depending on the sample size, item discrimination is not expected to be excessively high. Also, each item difficulty index ($b_i$ , $i$ = 1, 2, …, 22) is on the same metric as the latent variable ($\theta$) with value range over the set of real numbers, and a practical range between -2 and 2 to avoid too easy or too tricky items on the test (DeMars, 2010). It gives information on the amount of the latent variable ($\theta$) at which 50% of the students will get a correct score on each item. In as much as there seems to be no specific range of values to ascertain good difficulty index, empirical evidence has supported retaining items of the middle index of difficulty on the test (Ebel & Frisbie, 1991).

In line with unidimensionality assumption of IRT models (DeMars, 2010), a one-factor model of the PKMT with its 22 questions hypothesised to measure a common construct is evaluated using mean and variance adjusted unweighted least squares estimator with theta parametrisation (ULSMV-Theta). ULSMV-Theta is used because of its satisfactory performance and precision in estimating model parameters for a dichotomously scored IRT modelling in Mplus (Paek, Cui, Ozturk-Gubes & Yang, 2018). The model fit is assessed using multiple criteria. For an appropriate fit, we follow the recommendations of the ratio of chi-square value to the degree of freedom of less than 3coupled with a root mean square error of approximation (RMSEA) of less than .06 with non-significant 90% confidence interval (Brown, 2015), comparative fit and Tucker-Lewis indices (CFI and TLI) of greater than or close to .90 (Bentler, 1990). Further, we look at the significant level or otherwise of the factor loading of each of the items on the test. This is necessary to determine the contribution of these items to the test and to estimate each item reliability using standardised R-square values.

The qualitative method of data analysis takes the form of a cognitive interview. This interview was conducted to further probe and to determine the most likely reasons why some perceived too difficult questions based on the results of statistical analysis are not answered correctly. We rely on students' experience that voluntarily consented to take part in the interviews. In addition to the general consent to take part in the research project, special consent was requested from each student before the interview to audio record individual's utterances. The semi-structured cognitive interview was individually conducted in Norwegian using some leading questions with samples as follow:

> *If you, please take a look at this task, do you think this is a task you would have mastered?*
>
> *Do you have any idea on how to solve that one? Is it clear what they ask?*
>
> *What do you think is the reason why many students got this question incorrect?*
>
> *You get some calculations there, and you only have paper and pencil accessible, do you think that a calculator had been necessary for some students on this task?*

A total of seven students were interviewed, including six males and one female. Each interview lasted about 15 minutes, and the collected data were transcribed and translated into English. Selected results from these interviews are presented in the next section.

## 4. Results and Discussion

### 4.1. Results of Quantitative Analyses

Results from the analysis of a one-factor model of the PKMT with its 22 questions hypothesised to measure a common construct of students' prior mathematics knowledge are presented. Descriptive statistics of the analysed data as well as some initial parameters are shown in Table 1. The table shows the number of correct and incorrect responses of each item on the test, including the respective standardised factor loadings, R-square values, and the p-values.

| Question | Number of correct responses | Number of incorrect responses | Factor loading | p-value | R-square | p-value |
|---|---|---|---|---|---|---|
| **1A** | 173 | 28 | .388 | .001 | .151 | .091 |
| **1B** | 123 | 78 | .472 | < .001 | .223 | .006 |
| **1C** | 103 | 98 | .444 | < .001 | .197 | .006 |
| **2A** | 151 | 50 | .454 | < .001 | .206 | .017 |
| **2B** | 112 | 89 | .562 | < .001 | .316 | < .001 |
| **2C** | 85 | 116 | .537 | < .001 | .288 | < .001 |
| **3** | 64 | 137 | .530 | < .001 | .281 | < .001 |
| **4** | 116 | 85 | .557 | < .001 | .310 | < .001 |
| **5** | 143 | 58 | .549 | < .001 | .301 | < .001 |
| **6** | 128 | 73 | .660 | < .001 | .436 | < .001 |
| **7** | 113 | 88 | .679 | < .001 | .461 | < .001 |
| **8** | 133 | 68 | .804 | < .001 | .646 | < .001 |
| **9A** | 48 | 153 | .169 | .078 | .029 | .379 |
| **9B** | 19 | 182 | .336 | < .001 | .113 | .034 |
| **10** | 13 | 188 | .553 | < .001 | .306 | .001 |
| **11A** | 176 | 25 | .891 | < .001 | .794 | .003 |
| **11B** | 60 | 141 | .350 | < .001 | .123 | .033 |
| **12** | 82 | 119 | .734 | < .001 | .539 | < .001 |
| **13** | 58 | 143 | .604 | < .001 | .365 | < .001 |
| **14** | 86 | 115 | .618 | < .001 | .382 | < .001 |
| **15** | 125 | 76 | .736 | < .001 | .542 | < .001 |
| **16** | 67 | 134 | .695 | < .001 | .483 | < .001 |

Table 1. Descriptive statistics of the 22-item PKMT

The results presented in Table 1 reveal that Question 11A of the PKMT has the highest number of correct responses with 176 students got it correctly while Question 10 has the least number of correct responses with only 13 students got it correctly. All the item factor loadings are significant except for Question 9A, which has an insignificant factor loading of .169 ($p = .078$). These factor loadings reflect the strength at which each of the questions of the PKMT measures the purported prior mathematics knowledge the instrument is designed to measure. Thus, from this initial analysis, one can deduce that Question 9A has little or no substantial contribution to the instrument. Further, upon squaring each of these standardised factor loadings, a measure of variability (R-square) and reliability of each question on the PKMT was established. For instance, 31.6% and 79.5% variances of Question 2B and Question 11A, respectively, are explained by the latent construct of students' competence on the PKMT. And that these questions are reliable with significant reliability coefficients of .316 and .795, respectively. On the other hand, questions 1A, 2A, 9A, 9B and 11B have non-significant reliability coefficients of .151, .206, .029, .113, and .123, respectively, at $\alpha = .01$ level of significance. The reliability of the whole test was found to be .92 using a latent variable approach described in (Raykov, Dimitrov & Asparouhov,

2010; Raykov & Marcoulides, 2016). The goodness of fits statistics from the analysis of the one-factor PKMT model are presented in Table 2.

| Model fit statistics | Values |
|---|---|
| ***Chi-square ($\chi^2$)*** | |
| *Value* | 272.892 |
| *df* | 209 |
| $\chi^2/df$ | 1.306 |
| *p-value* | .002 |
| ***CFI/TLI*** | |
| *CFI* | .903 |
| *TLI* | .893 |
| ***RMSEA*** | |
| *Estimate* | .039 |
| *90 per cent C.I.* | .025 .051 |
| *Probability RMSEA <= .05* | .927 |

Table 2. Selected goodness of fit indices of the one-factor PKMT model

The results presented in Table 2 show an appropriate fit of the evaluated one-factor model of the PKMT. The chi-square statistic seems a bit high and significant (p=.002). However, its ratio to the degree of freedom is less the recommended value of 3 for an acceptable model fit. Both the CFI and the TLI values are within the recommended values of an acceptable model fit (Bentler, 1990). The RMSEA is excellent with its value within the 90 per cent confidence interval, and its probability is not significant. This non-significant RMSEA probability shows that the model demonstrates a close fit of the data and that the hypothesis of not-close fit should be rejected (MacCallum, Browne & Sugawara, 1996). Thus, the overall fit statistics confirm that the hypothesised one-factor construct of prior mathematics knowledge exposed by the 22 questions is supported by empirical evidence. After establishing the model fit of the PKMT, we now turn to its item quality as explicated by item response theory parametrisation. The ensuing results on item discrimination and difficulty indices of each item on the PKMT as well as their respective p-values are presented in Table 3.

The results presented in Table 3 show that all the questions on the PKMT have acceptable item discrimination indices except for Question 9A ($a_{9A} = 0.172$, $p = .087$) and Question 11A ($a_{11A} = 1.958$, $p = .223$) which demonstrate too weak and too strong discriminations, respectively, among the students. The inference can be drawn from the non-significant estimates of the discrimination indices of these two questions. According to the classifications of item discrimination index by Ebel and Frisbie (1991), it can be inferred that our empirical evidence supports the removal of Question 9A and Question 11A from the test, Questions 9B and 11B have appropriate discriminating indices but can be improved upon, and all other questions have good discrimination indices. Further, it is also revealed in Table 3 that some questions demonstrate appropriate difficulty. At the same time, some questions demonstrate excessive item difficulty (i.e. too difficult questions), and other questions demonstrate weak difficulty (i.e. easy questions). For instance, questions 1C, 2B, 2C, 4, 7, 9A, 12, and 14 demonstrate appropriate difficulty with the non-significant estimates ($p > .01$) of their respective difficulty indices. Also, questions 1A, 1B, 2A, 5, 6, 8, 11A, and 15 are relatively easy questions depending on the absolute magnitude of their estimates while other questions, e.g. questions 3, 9B, 10, and 11B are difficult questions. Selected results of why students perceived some of these questions difficult are presented in the next section.

| Question | Item discrimination | p-value | Item difficulty | p-value |
|---|---|---|---|---|
| **1A** | 0.421 | .004 | -2.795 | .002 |
| **1B** | 0.536 | < .001 | -0.602 | .006 |
| **1C** | 0.495 | < .001 | -0.070 | .725 |
| **2A** | 0.510 | < .001 | -1.494 | < .001 |
| **2B** | 0.679 | < .001 | -0.256 | .112 |
| **2C** | 0.637 | < .001 | 0.362 | .040 |
| **3** | 0.626 | < .001 | 0.890 | < .001 |
| **4** | 0.670 | < .001 | -0.350 | .034 |
| **5** | 0.657 | < .001 | -1.016 | < .001 |
| **6** | 0.877 | < .001 | -0.531 | < .001 |
| **7** | 0.924 | < .001 | -0.231 | .080 |
| **8** | 1.354 | < .001 | -0.518 | < .001 |
| **9A** | 0.172 | .087 | 4.198 | .090 |
| **9B** | 0.357 | < .001 | 3.908 | < .001 |
| **10** | 0.664 | < .001 | 2.741 | < .001 |
| **11A** | 1.958 | .223 | -1.295 | < .001 |
| **11B** | 0.374 | < .001 | 1.509 | .001 |
| **12** | 1.081 | < .001 | 0.317 | .012 |
| **13** | 0.758 | < .001 | 0.923 | < .001 |
| **14** | 0.786 | < .001 | 0.294 | .049 |
| **15** | 1.086 | < .001 | -0.422 | .001 |
| **16** | 0.965 | < .001 | 0.620 | < .001 |

Table 3. IRT parameterisation of the PKMT

### 4.2. Results of Cognitive Interviews

To further probe why some questions are perceived difficult by the students, we interviewed some students to hear their views and suggestions for the improvement of such difficult questions. Results from the transcripts of interviews for Question 10 (this is a word-problem type question that requires the students to manipulate some percentages and give the final answer in decimal number) show that some students find it challenging to understand the question because of its practical and word-problem nature. Some of the reasons stated for getting the question incorrect by most students are lack of proper understanding of the question, text misreading, and unavailability of calculators. The students also think that provision of calculators during the test administration could improve their performance on such difficult tasks. For the reason that they are used to working on mathematical tasks with calculators lately, as mentioned by one of the students "*I would have thought about this for a while, I guess I had to because we are so used to use the calculator all the time*".

Similarly, when the interviewer asked the following questions about Question 9: What is difficult here? What makes this a bit difficult? If you could try to describe in words what makes it difficult to understand? Note: Question 9 is a word-problem type that requires the students to manipulate the purchase of oranges and bananas in kilogrammes using letters rather than numbers. One of the interviewees responded with the following answer:

> *No, it is more. I think it was difficult to understand. That the **a** stands for, I am more like I do not manage to deal with practical tasks after I began doing theoretical tasks […] This is a typical example of what I find difficult, that **a** is how many kilograms of oranges that you buy, and **b** is for bananas. What is 10**a** plus 15**b**? And then I become, like – actually I think I could have solved it if I had more time. If I had thought more about it. It is not how to solve it; it is more like I put a lot more energy into solving this task than this task [she points to task 3 which is on calculating the volume of a compound figure] because it is too much text and I become stressed, and I think back to the practical math that I had and that I did not like.*

It can be deduced from the excerpts of interview transcriptions for Question 9 that some students could not solve the problem correctly because of their inadequate reading comprehension, interpretations, and improper understanding of the word-problem task. Meanwhile, of the six difficult questions (3, 9B, 10, 11B, 13 and 16) identified in Table 3, only questions 3 and 11B are not posed in word problems. Thus, it can be inferred that the challenge with our students lies on their improper grasping of word-problem tasks which could stem from their preference for other types of mathematical tasks as evident in one of the student's response during the interview "*It is not how to solve it, it is more like I put a lot more energy into solving this task than this task* [she points to task 3 which is on calculating the volume of a compound figure] *because it is too much text and I become stressed, and I think back to the practical math that I had and that I did not like"*. This finding conforms to the global trend of students' perceived difficulty of mathematical word-problem tasks at elementary, secondary and university levels (e.g., Vilenius-Tuohimaa, Aunola & Nurmi, 2008; Zheng, Swanson & Marcoulides, 2011).

## 5. Conclusions

Prior mathematics knowledge of students has been identified as instrumental to the learning outcomes of current materials. Both theoretical and empirical evidence has been documented to support this claim (Bandura, 1997; Zakariya, 2016). However, proper assessment of students' prior mathematics knowledge has been beset with inconsistency in the available numerous measuring instruments and lack of validation studies. Attempts are made in the present study to validate a national test of prior mathematics knowledge of university students in Norway using mixed methods research design. The design involves the use of item response theory to provide psychometric properties of the test and cognitive interviews to probe plausible reasons why students find some questions challenging.

The findings of the present study provide empirical evidence for the construct validity of the Norwegian prior knowledge of mathematics test. In particular, our evaluation of a one-factor model shows that the test is measuring just a single latent variable (i.e. prior mathematics knowledge of students) that it is purported to measure. Further, it is also found that out of the 22 questions on the test only questions 1A, 2A, 9A, 9B and 11B demonstrate lack of acceptable reliability coefficients. However, the reliability coefficient of the whole test using latent variable approach is found to be very high (.92) which proves high internal consistency of the items on the test (Raykov et al., 2010). The latent variable approach is used to compute the reliability coefficient of PKMT because of its reported excellent performance over the popular Cronbach's alpha and Kuder-Richardson formula 20 (e.g., Raykov et al., 2010; Raykov & Marcoulides, 2016). In as much as most of the reviewed literature in the present study (e.g., Hailikari et al., 2008; Newman-Ford et al., 2009) do not report reliability coefficients of their measures of prior academic knowledge, the reliability coefficient of the PKMT is higher than the one reported by Lee and Chen (2009) but slightly lower than the Kuder-Richardson coefficient reported by Casillas et al. (2012).

The findings of the present study also show that questions on the PKMT are at different levels of difficulty and variant discriminations between students of low and high competence in the prior mathematics knowledge test. These findings have several implications on the validity and reliability of aggregate scores of the test and other analyses (e.g. means comparisons between universities and previous years) usually presented by the Norwegian Mathematical Council. For instance, the assignment of a score of 1 point to an easy and poorly discriminating item, e.g., Question 11A and to a challenging and good discriminating item, e.g., Question 10 may bias the aggregate scores of students with low ability upward on the test and reduce the aggregate scores of highly competent students. This kind of bias in aggregate scores is a threat to the validity of the test and a typical disadvantage of using classical test theory approach in scoring tests (DeMars, 2010). Thus, we urge the Norwegian Mathematical Council to use item response theory which can incorporate the test item difficulty and discrimination indices in the scoring process such that more valid aggregate scores can be obtained. Moreover, of course, a more reliable mean score comparison can be made. Further, compelling evidence is also provided in the findings of the present study that suggests the removal of or at least improvement in item wordings and presentation of questions 9A, 9B, 11A and 11B on the test.

Moreover, six out of the 22 questions of PKMT are also found to be very difficult for students to answer correctly. Empirical evidence from cognitive interviews of some students who took part in the test reveals potential reasons why these questions are perceived difficult. Some of the ascribed causes of poor performance on these questions are lack of proper understanding of the question, text misreading, improper grasping of word-problem mathematical tasks, and unavailability of calculators. Given that low performance on word-problem tasks is not peculiar to Norwegian engineering students (e.g., Vilenius-Tuohimaa et al., 2008), we recommend innovative teaching and learning strategies to alleviate these problems. Such strategies can be the use of modelling activities, problem-based learning, and so on (Greer, 1997; Zakariya, Ibrahim & Adisa, 2016) to foster understanding and interpretation of word-problem mathematical tasks. The Norwegian Mathematical Council may also consider the introduction of calculators in subsequent PKMT administrations. Finally, this section is concluded by acknowledging some strengths and potential weaknesses of the findings of this study.

## 6. Strengths and Potential Weaknesses of this Study

A strength of this study lies in the use of explanatory sequential mixed-methods approach to data analysis (Bryman, 2016) that involves a robust quantitative analysis procedure in terms of an IRT followed by some cognitive interviews. The interviews avail us the opportunity to look at the data beyond statistical analyses and provide a more elaborate description of the phenomenon. Another strength of this study encompasses a relatively large data set of 201 engineering students used in the present study. The large sample involved is a potential for generalisation of our findings, especially now that such large-scale study is scarce in mathematics education research. However, a potential limitation of the present study could stem from a lack of external validity of the PKMT. There was no independently measured variable such as students' grades, and grade point average through which the predictive validity of the PKMT can be confirmed. We did not investigate the content validity of the test items as we lack the permission to do so. Instead, our findings only provide evidence for its psychometric property. Also, the restriction of the sample of the study to a Norwegian university and only engineering students might, in a way, limit the generalisation of our findings. Thus, we recommend the replications of the present study in more substantial and more diverse university student populations. Despite these limitations, our study does provide potential cues on the construct validity, reliability, and item quality of the PKMT which will be useful to Norwegian Mathematical Council, researchers, and other stakeholders in mathematics education. The methodology adopted in the present study can also be applied in other national contexts to investigate the validity of their measures.

## Acknowledgements

The authors appreciate the kind gesture of the Norwegian National Mathematical Council for permitting us to use their test.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest for the research, authorship, and publication of this article.

## Funding

The authors received no financial support for the research and authorship of this article. Meanwhile, we appreciate the University of Agder library for funding the article processing charge of this article.

## References

Aluko, R.O., Daniel, E.I., Oshodi, O.S., Aigbavboa, C.O., & Abisuga, A.O. (2018). Towards reliable prediction of academic performance of architecture students using data mining techniques. *Journal of Engineering, Design and Technology*, 16(3), 385-397. https://doi.org/10.1108/jedt-08-2017-0081

Ayán, M.N.R., & García, M.T.C. (2008). Prediction of university students' academic achievement by linear and logistic models. *The Spanish Journal of Psychology*, 11(1), 275-288. https://doi.org/10.1017/s1138741600004315

Bandura, A. (1997). S*elf-efficacy: The exercise of control*. New York: W.H. Freeman and Company.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Bloom, B.S. (1976). *Human characteristics and school learning*. New York: McGraw-Hill.

Brown, T.A. (2015). *Confirmatory factor analysis for applied research*. New York, London: The Guilford Press.

Bryman, A. (2016). *Social research methods*. Oxford, United Kingdom: Oxford University Press.

Casillas, A., Robbins, S., Allen, J., Kuo, Y.L., Hanson, M.A., & Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. *Journal of Educational Psychology*, 104(2), 407-420. https://doi.org/10.1037/a0027180

DeMars, C. (2010). *Item response theory: Understanding statistics measurement*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

Dochy, F. (1996). Assessment of domain-specific and domain-transcending prior knowledge: Entry assessment and the use of profile analysis. In Birenbaum M., & Dochy, F. (Eds.), *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge. Evaluation in Education and Human Services* (42, 227-264).Dordrecht: Springer. https://doi.org/10.1007/978-94-011-0657-3_9

Dochy, F., De Rijdt, C., & Dyck, W. (2002). Cognitive prerequisites and learning. How far have we progressed since Bloom? Implications for educational practice and teaching. *Active learning in higher education*, 3(3), 265-284. https://doi.org/10.1177/1469787402003003006

Duff, A. (2004). Understanding academic performance and progression of first-year accounting and business economics undergraduates: The role of approaches to learning and prior academic achievement. *Accounting Education*, 13(4), 409-430. https://doi.org/10.1080/0963928042000306800

Ebel, R., & Frisbie, D.A. (1991). *Essentials of educational measurement*. New Delhi: Prentice - Hall of India.

Engerman, K., & Bailey, U.J.O. (2006). Family decision-making style, peer group affiliation and prior academic achievement as predictor of the academic achievement of African American students. *The Journal of Negro Education*, 75(3), 443-457.

Greer, B. (1997). Modelling reality in mathematics classrooms: the case of word problems. *Learning and Instruction*, 7(4), 293-307. https://doi.org/10.1016/S0959-4752(97)00006-6

Hailikari, T., Nevgi, A., & Komulainen, E. (2008). Academic self-beliefs and prior knowledge as predictors of student achievement in Mathematics: A structural model. *Educational Psychology*, 28(1), 59-71. https://doi.org/10.1080/01443410701413753

Lee, C.Y., & Chen, M.P. (2009). A computer game as a context for non-routine mathematical problem solving: the effects of type of question prompt and level of prior knowledge. *Computers & Education*, 52(3), 530-542. https://doi.org/10.1016/j.compedu.2008.10.008

MacCallum, R.C., Browne, M.W., & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods,* 1(2), 130-149. https://doi.org/10.1037/1082-989X.1.2.130

Martin, A.J., Wilson, R., Liem, G.A.D., & Ginns, P. (2016). Academic momentum at university/college: Exploring the roles of prior learning, life experience, and ongoing performance in academic achievement across time. *The Journal of Higher Education*, 84(5), 640-674. https://doi.org/10.1353/jhe.2013.0029

Marton, F., & Booth, S.A. (1997). *Learning and awareness*. Hillsdale, NJ: Lawrence Erlbaum.

Newman-Ford, L., Lloyd, S., & Thomas, S. (2009). An investigation in the effects of gender, prior academic achievement, place of residence, age and attendance on first-year undergraduate attainment. *Journal of Applied Research in Higher Education*, 1(1), 14-28. https://doi.org/10.1108/17581184200800002

Opstad, L., Bonesrønning, H., & Fallan, L. (2017). Tar vi opp de rette studentene ved økonomisk-administrative studier? *Samfunnsøkonomen,* 1, 21-29.

Paek, I., Cui, M., Ozturk-Gubes, N., & Yang, Y. (2018). Estimation of an IRT model by Mplus for dichotomously scored responses under different estimation methods. *Educational and Psychological Measurement*, 78(4), 569-588. https://doi.org/10.1177/0013164417715738

Raykov, T., Dimitrov, D.M., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 17(2), 265-279. https://doi.org/10.1080/10705511003659417

Raykov, T., & Marcoulides, G.A. (2016). Scale reliability evaluation under multiple assumption violations. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(2), 302-313. https://doi.org/10.1080/10705511.2014.938597

Richardson, M., & Abraham, C. (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

Vilenius-Tuohimaa, P.M., Aunola, K., & Nurmi, J.E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology*, 28(4), 409-426. https://doi.org/10.1080/01443410701708228

Zakariya, Y.F. (2016). Investigating validity of Math 105 as prerequisite to Math 201 among undergraduate students, Nigeria. *International Journal of Instruction*, 9(1), 107-118. https://doi.org/10.12973/iji.2016.919a

Zakariya, Y.F., Ibrahim, M.O., & Adisa, L.O. (2016). Impacts of problem-based learning on performance and retention in mathematics among junior secondary school students in Sabon-Gari area of Kaduna state. *International Journal for Innovative Research in Multidisciplinary Field*, 2(9), 42-47.

Zheng, X., Swanson, H.L., & Marcoulides, G.A. (2011). Working memory components as predictors of children's mathematical word problem solving. *Journal of Experimental Child Psychology*, 110(4), 481-498. https://doi.org/10.1016/j.jecp.2011.06.001

Published by OmniaScience (www.omniascience.com)

Journal of Technology and Science Education, 2020 (www.jotse.org)