Received: 16 July 2020 | Revised: 6 November 2020 | Accepted: 23 January 2021

DOI: 10.1111/jan.14799

RESEARCH METHODOLOGY: INSTRUMENT DEVELOPMENT

Development and psychometric testing of Reasoning Skills test for nursing student selection: An item response theory approach

Jonna Vierula[1] | Kirsi Talman[1] | Maija Hupli[1] | Eero Laakkonen[2] | Janne Engblom[3] | Elina Haavisto[4]

[1]Department of Nursing Science, University of Turku, Turku, Finland

[2]Department of Teacher Education, University of Turku, Turku, Finland

[3]Department of Accounting and Finance, Turku School of Economics, University of Turku, Turku, Finland

[4]Department of Nursing Science, Hospital District of Satakunta, University of Turku, Turku, Finland

Correspondence

Jonna Vierula, Department of Nursing Science, University of Turku, Turku, Finland.
Email: johevi@utu.fi

Funding information

The study was supported by a doctoral grant from the not-for-profit sector of The Finnish Nursing Education Foundation and The Finnish Association of Nursing Research.

ABSTRACT

Aims: To develop and psychometrically test the Reasoning Skills (ReSki) test assessing undergraduate nursing applicants' reasoning skills for student selection purposes.

Design: A methodological cross-sectional design was applied for the psychometric testing.

Methods: The ReSki test was developed as part of a wider electronic entrance examination. The ReSki test included a case followed by three question sections assessing nursing applicants' reasoning skills according to the reasoning process. Item response theory was used for psychometric testing to assess item discrimination, difficulty and pseudoguessing parameters. The ReSki test was taken by 1056 nursing applicants in six Finnish Universities of Applied Sciences (28 May 2019).

Results: In the development process, the expert evaluations indicated acceptable content validity. In the psychometric testing, the test reliability was supported by item variance, the theoretical structure was supported by the correlation coefficients and the applicant mean performance supported an acceptable overall test difficulty. The item response theory indicated variance between the items' difficulty and discrimination ranges. However, most of the wrong items failed at being functional distractors.

Conclusion: The ReSki test is a new and valid objective assessment of undergraduate nursing applicants' reasoning skills. The item response theory provided item-level information that can be used for further development of the test, especially related to the revisions needed for the distractor items to achieve the desired level of difficulty.

Impact: *What problem did the study address?* The assessment of nursing applicants' reasoning skills is suggested, but there is a lack of admission tools. *What were the main findings?* The results provided support for the reliability and validity of the ReSki test. Item response theory indicated the need for further item-level improvement. *Where and on whom will the research have an impact?* The results may benefit higher education institutions and researchers when developing a test and/or student selection processes.

KEYWORDS
instrument development, item response theory, nursing, psychometric testing, reasoning, skills, student selection, test development, undergraduate nursing education

1 | INTRODUCTION

Reasoning skills have been identified as important to be assessed in nursing student selection (Haavisto et al., 2019; Vierula, Haavisto, et al., 2020; Vierula, Hupli, et al., 2020). Cognitive skills such as reasoning are vital in nursing (Kajander-Unkuri et al. 2013) and important to assess already in the selection phase to ensure an applicant's academic progress and the attainment of professional qualifications (Perkins et al., 2013; Vierula, Haavisto, et al., 2020; Vierula, Hupli, et al., 2020). Reasoning skills are essential in theoretical (McNelis et al., 2010) and clinical (Timer & Clauson, 2011) studies. Good reasoning skills enable solid decision-making improving patient safety. The importance of reasoning skills is emphasized in contemporary healthcare settings that are cognitively demanding, characterized by the increasing use of technology and patients with complex health problems (Levett-Jones et al., 2010). Although the assessment of reasoning skills has been previously suggested for the selection phase, there is a lack of admission tools and a need for operationalization of the concept (Haavisto et al., 2019; Schmidt & MacWilliams, 2011; Vierula, Haavisto, et al., 2020; Vierula, Hupli, et al., 2020). In this study, the Reasoning Skills (ReSki) test was developed and tested to assess undergraduate nursing applicants' reasoning skills.

Since 2000, the number of nursing students has increased in the majority of the Organisation for Economic Co-operation and Development countries (OECD, 2019). In 2017, there were 121,000 nursing graduates in the European Union (Eurostat, 2019), reflecting the large number of applications processed yearly. Higher education institutions (HEIs) must look for suitable applicants who are able to fulfil the competence requirements of nursing (Schmidt & MacWilliams, 2011) and deliver safe care as future professionals (Francis, 2013; Shulruf et al., 2018). Recently, many countries have been developing nursing/medical student selection processes, aiming to determine exactly what to assess, how to assess it and whom to select (Haavisto et al., 2019; MacDuff et al., 2016; Shulruf et al., 2018). Moreover, HEIs are responsible for choosing their applicants fairly (Shulruf et al., 2018) by employing valid and reliable admission tools (Perkins et al., 2013). Student selection is part of high-stakes testing, having important consequences both for the test-taker and HEI (National Council on Measurement in Education [NCME], 2017). Developing a valid test is necessary for the equal treatment of applicants. Validity of an admission test is the degree to which all the accumulated evidence supports the intended interpretation of test scores for proposed use, including the evidence based on test content, response processes, internal structure, relations to other variables and consequences of testing (American Educational Research Association [AERA], American Psychological Association [APA], & NCME, 2014). Traditionally, the psychometric properties of tests/exams are reported using classical test theory (CTT). However, an item response theory (IRT) approach is increasingly recommended for assessing the validity of measurement scales in nursing (Tavakol et al., 2014; Yang & Kao, 2014). In this study, the psychometric properties were examined using the IRT approach.

2 | BACKGROUND

Reasoning, decision-making, problem-solving, clinical judgement and critical thinking are all concepts used interchangeably to describe cognitive skills (Carbogim et al., 2016; Simmons, 2010). All these concepts include elements of both the thinking process and its outcome, but reasoning focuses on the thinking process, whereas making decisions and judgements and solving problems refer more to an end point of the reasoning process (Simmons, 2010). Critical thinking is a facilitator of clinical reasoning (Alfaro-LeFevre, 2013; Facione, 1990), involving knowledge, experiences, dispositions (attitudes or habits of mind) and intellectual abilities (Carbogim et al., 2016) being more than learned skills (Facione et al., 1994). Reasoning is a cognitive skill and a generic term, defined as a cognitive process directed towards forming conclusions, judgements or inferences based on facts or premises (Merriam Webster Dictionary, 2020; Simmons, 2010). In clinical settings, reasoning refers to a complex cognitive process, which uses formal and informal thinking strategies to gather and analyse patient information, evaluate the significance of this information and weigh alternative actions (Levett-Jones et al., 2010; Simmons, 2010). According to Vierula, Hupli, et al. (2020), nursing applicants' reasoning skills should include the ability to identify essential information, process the information and use the information to make decisions. In this study, reasoning skills were understood to be the cognitive skills, abilities, readiness and aptitudes required to gain entry to a nursing programme.

Vierula, Haavisto, et al. (2020) identified that nursing applicants' reasoning skills have been measured with the attributes of critical thinking using the Health Sciences Reasoning Test and the Watson–Glaser Critical Thinking Appraisal. However, the ambiguous nature of the concept and challenges in measurement have been pointed out (Carbogim et al., 2016; Zuriguel Pérez et al., 2014). Decision-making, problem-solving and logics have been assessed using other onsite selection methods than standardized tests (e.g., an interview; Vierula, Haavisto, et al., 2020). In medical schools, applicants' reasoning skills have been measured with the University Clinical Aptitude Test (UCAT, 2020). Although some of the existing tests measure applicants' reasoning skills, the operationalization of the concept varies, leaving the question of what exactly to assess (Vierula, Haavisto, et al., 2020). The assessment should focus on cognitive domains that comprehensively reflect the requirements of the professional education and identify applicants who are most likely to succeed (Schmidt & MacWilliams, 2011; Vierula, Haavisto, et al., 2020; Vierula, Hupli, et al., 2020).

3 | THE STUDY

3.1 | Aim

The aim of the study was to develop and test the psychometric properties of the ReSki test for assessing undergraduate nursing applicants' reasoning skills for student selection purposes. The study is part of the Reforming Student Selection in Nursing Education (ReSSNE) project developing the undergraduate (bachelor-level) nursing student selection processes in Finland (Haavisto et al., 2019).

3.2 | Methodology

The ReSki test was developed following the scale development procedure applying scoping reviews, focus group interviews, expert evaluations and pilot tests (DeVellis, 2017; Figure 1). A methodological cross-sectional design with the IRT approach was applied to test the psychometric properties of the developed test. IRT is based on mathematical equations explaining the relationship between test-taker ability and the probability of a correct/incorrect item response using a nonlinear monotonic function (Hays et al., 2000). The focus of the assessment is shifted to the individual items, thus enabling the

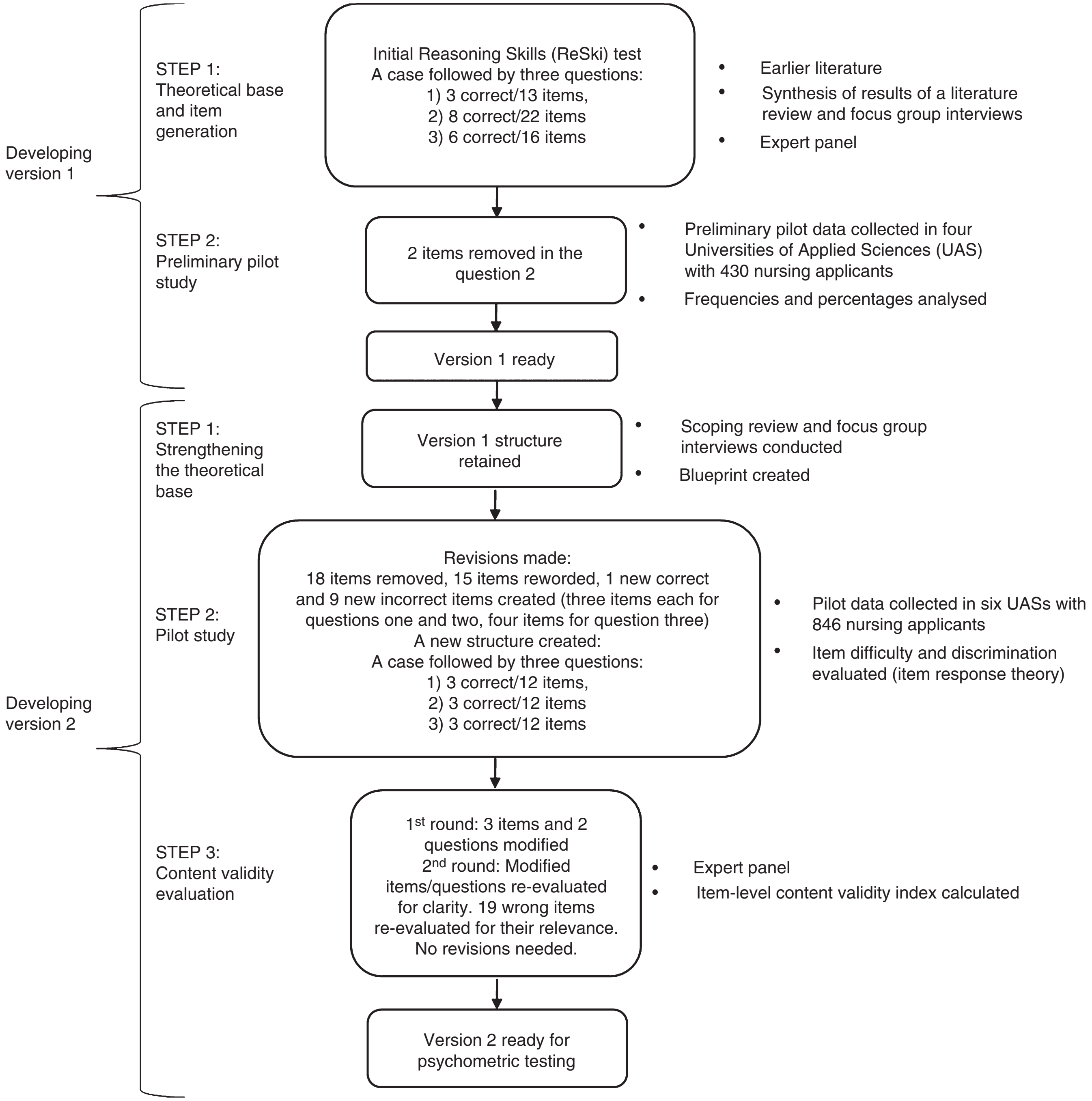
Developing version 1
STEP 1: Theoretical base and item generation
Initial Reasoning Skills (ReSki) test
A case followed by three questions:
1) 3 correct/13 items,
2) 8 correct/22 items
3) 6 correct/16 items
• Earlier literature
• Synthesis of results of a literature review and focus group interviews
• Expert panel
STEP 2: Preliminary pilot study
2 items removed in the question 2
• Preliminary pilot data collected in four Universities of Applied Sciences (UAS) with 430 nursing applicants
• Frequencies and percentages analysed
Version 1 ready
Developing version 2
STEP 1: Strengthening the theoretical base
Version 1 structure retained
• Scoping review and focus group interviews conducted
• Blueprint created
STEP 2: Pilot study
Revisions made:
18 items removed, 15 items reworded, 1 new correct and 9 new incorrect items created (three items each for questions one and two, four items for question three)
A new structure created:
A case followed by three questions:
1) 3 correct/12 items,
2) 3 correct/12 items
3) 3 correct/12 items
• Pilot data collected in six UASs with 846 nursing applicants
• Item difficulty and discrimination evaluated (item response theory)
STEP 3: Content validity evaluation
1st round: 3 items and 2 questions modified
2nd round: Modified items/questions re-evaluated for clarity. 19 wrong items re-evaluated for their relevance. No revisions needed.
• Expert panel
• Item-level content validity index calculated
Version 2 ready for psychometric testing

FIGURE 1 Development process of the Reasoning Skills (ReSki) test

evaluation of item parameters, such as discrimination, difficulty and pseudoguessing (the possibility of guessing/false positive; DeVellis, 2017; Tavakol et al., 2014). Moreover, IRT is considered useful for nurse educators and researchers to obtain a greater understanding of the interaction between test-taker ability and item parameters to monitor and improve the quality of an assessment (Dimitrov & Shelestak, 2003; Tavakol et al., 2014).

The two most well-known IRT models are the one-parameter (difficulty) logistic (1PL/Rasch analysis) model and the two-parameter (difficulty and discrimination) logistic (2PL) model (Sulis & Toland, 2017). Additionally, a pseudoguessing parameter is possible to examine by using IRT (DeVellis, 2017; Tavakol et al., 2014). In IRT, the relationship between test-taker ability and item parameters is transformed mathematically by using natural logarithms in an interval scale, resulting in a visual S-shaped logistic curve: the item characteristic curve (ICC; DeVellis, 2017; Tavakol et al., 2014; Figure 2). In ICCs, the *Y*-axis is the probability of a correct response. The *X*-axis is the ability increasing from right to left typically ranging from ±3. Discrimination describes the relationship of performance for an item relative to performance on the full test. Difficulty reflects the point on the scale where the likelihood of a correct response is 50%. The slope of the ICC reveals if an item discriminates among weak test-takers: The steeper the slope, the greater the discrimination. The shift of the ICC reveals the difficulty of an item: An easy item shifts the curve to the left along the horizontal axis (test-taker ability), whereas a difficult item shifts the curve to the right. The pseudoguessing parameter is revealed by looking at the starting point of an ICC along the vertical axis ranging from 0 to 1: the higher the starting point of an ICC, the higher the possibility of guessing.

3.3 | Development of the ReSki test

The ReSki test was developed during the period of 2015–2019 for student selection purposes as part of a wider electronic entrance examination measuring undergraduate nursing applicants' language skills, mathematical skills, emotional intelligence and certainty of career choice (Haavisto et al., 2019). The first version of the ReSki test was developed in two steps. The second version was developed in three steps (Figure 1).

3.3.1 | Development of version 1

Step 1: Theoretical base and item generation

The first version of the ReSki test was developed simultaneously with the other domains of the new entrance examination. The ReSki test was based on previous studies (Levett-Jones et al., 2010; Simmons, 2010), a synthesis of a literature review (*N* = 13) and focus group interviews (*N* = 27) with nurse educators, directors, students and a nursing association representative (Haavisto et al., 2019; Figure 1). The test was structured on a case format, including a written case followed by three question sections, to assess nursing applicants' reasoning skills according to the reasoning process (Levett-Jones et al., 2010; Simmons, 2010): (1) collecting and processing information, (2) identifying the problem and (3) finding the solution (Figure S3). The case was related to overweight, a phenomenon that the applicants would be familiar with, but for which no previous knowledge in nursing would be required.

The number of items (answer options) per question was formulated on four bases. First, the items needed to follow the cues given in the case (Levett-Jones et al., 2010). Second, the wrong items needed to be functional distractors, meaning that they were tempting but incorrect (Malau-Aduli & Zimitat, 2012). Third, the number of items should reduce false positives by decreasing the possibility of guessing (Chiu & Camilli, 2013). Fourth, the number of initial items should be rather large because items are often reduced during the instrument testing (DeVellis, 2017). After the initial item generation, a panel of experts (*N* = 5), including nurse educators and researchers having expertise in reasoning skills, evaluated the test items to gain consensus (Graham, 2010; Polit & Beck, 2006). The number of evaluation rounds was not decided beforehand, and two evaluation rounds were undertaken. One item was modified, and three new items were generated for question two, resulting in this test structure: first question, 13 items in total (three correct); second

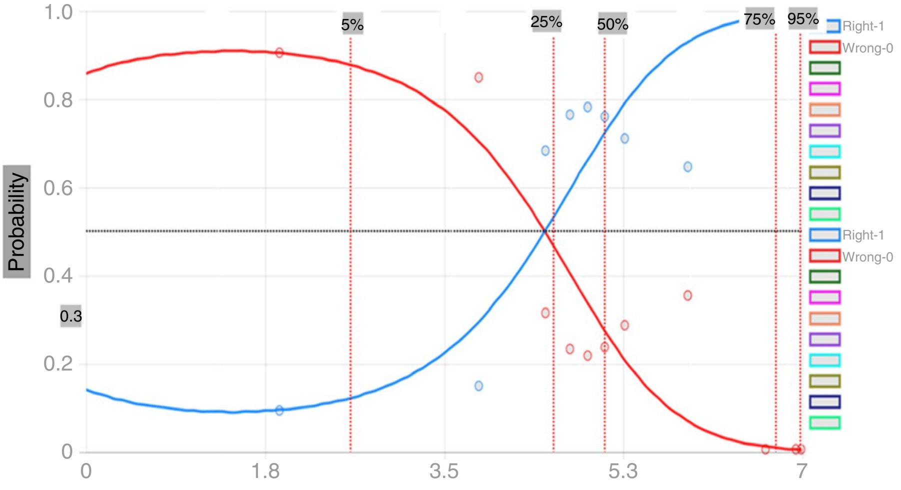

FIGURE 2 Example of an item characteristic curve. Pseudoguessing threshold and quantiles (5%, 25%, 50%, 75%, 95%) highlighted in grey [Colour figure can be viewed at wileyonlinelibrary.com]

question, 22 items in total (eight correct) and third question, 16 items in total (six correct; Figure 1; Figure S3). The total scores of the ReSki test were allocated as part of the overall entrance exam scores. Penalty scores were not used to avoid risk-taking strategies affecting the performance of the test-takers (Stenlund et al., 2017).

Step 2: Preliminary pilot study

A preliminary pilot study was conducted (3 November 2016) for the first application of the ReSki test (version 1; Maillard et al., 2017). The aim was to gain a preliminary understanding of the functionality and difficulty of the developed test. A total of 430 undergraduate nursing applicants from four Finnish universities of applied sciences (UASs) gave their consent for the study, took the test and completed a short feedback questionnaire. Frequencies and percentages were calculated for each item. Based on the feedback, only 23.2% of the applicants found the test difficult to very difficult. After the preliminary pilot, two dysfunctional distractors (>95% of the respondents answered correctly) were removed, and we determined the need to continue developing the test (Figure 1; Table 1).

3.3.2 | Development of version 2

Step 1: Strengthening the theoretical base of the test

A scoping review (*N* = 24; Vierula, Haavisto, et al., 2020) and focus group interviews (Vierula, Hupli, et al., 2020) with graduating nursing students and experts (educators, managers and researchers; *N* = 25) were conducted to strengthen the test's theoretical base (Figure 1). The aim was to identify the applicants' reasoning skills to be assessed to further establish and operationalize the concept (DeVellis, 2017). The development of test version 1 was conducted simultaneously with all the domains to be assessed in the student selection. Therefore, it was considered important to strengthen the theoretical base and gather more detailed information about reasoning skills for ensuring the development of a valid test.

As a result of the review, the lack of instruments focusing on the assessment of reasoning skills in the student selection was confirmed (Vierula, Haavisto, et al., 2020). According to the focus group results, the assessment should focus on the first steps of the reasoning process, meaning the assessment of cognitive process before

TABLE 1 Structure of the Reasoning Skills (ReSki) test versions 1 and 2

Measures	Questions and items	Scoring[a]	Format	Time to complete
ReSki test version 1				
Collecting and processing information	1st question: 3 correct items out of 13 items.	Correct answer = 0.5pt Wrong answer/don't know = 0pt	A case format electronic test	No specific time limitation is set. The test is a part of a wider electronic entrance examination that takes two and a half hours
Identifying the problem	2nd question: 8 correct items out of 20 items.	Correct answer = 0.25pt Wrong answer/don't know = 0pt		
Finding the solution	3rd question: 6 correct items out of 16 items.	Correct answer = 0.25pt Wrong answer/don't know = 0pt		
Reasoning Skills (ReSki) test version 2				
Collecting information	1st question: 3 correct items out of 12 items.	Correct answer = 0.5pt Wrong answer/don't know = 0pt	A case-format electronic test	No specific time limitation is set. The test is a part of a wider electronic entrance examination that takes two and a half hours
Processing information	2nd question: 3 correct items out of 12 items.	Correct answer = 0.5pt Wrong answer/don't know = 0pt		
Identifying the problem and establishing goals	3rd question: 3 correct items out of 12 items.	Correct answer = 0.5pt Wrong answer/don't know = 0pt		

[a] No penalty scores are given. The test scoring is part of a wider electronic entrance examination.

implementing the solution (Vierula, Hupli, et al., 2020). Next, a blueprint was created to overview the test and establish the emphasis of the measurement (Waltz et al., 2017). The structure of the initial test (version 1) was retained, but the blueprint supported renaming the skills to be measured as: (1) collecting information, (2) processing information and (3) identifying the problem and establishing goals (Table 1; Figure 1).

Step 2: Pilot study

The ReSki test (version 1) was piloted (31 October 2018) in six Finnish UASs with undergraduate nursing applicants who took the electronic entrance examination and gave their consent to participate in the study (Figure 1). The major aim of the pilot study was to test the level of discrimination and the difficulty of the generated items using IRT (DeVellis, 2017). The pilot sample (*N* = 846, population 12,421) represented typical Finnish population characteristics. Most of the applicants were female (*N* = 738, 87.6%), young adults (mean age 26.5 years, standard deviation [SD] 8.1, range 18–57) and 44.1% were first-time applicants.

Descriptive statistics were used to overview the data. The item proportions in each question were analysed with percentages and frequencies. Extremely easy items (>95% of the respondents chose the correct option or <1% chose the wrong option) were removed prior to the IRT analysis (6/49 items). The discrimination and difficulty level of the remaining 43 items were analysed with the 2PL model for each question with Stata 15.1 (StataCorp., 2017) and Mplus 8.1 (Muthén & Muthén, 1998–2012) statistical programs. The items were deleted if the difficulty estimates were below −2 referring to very easy (Baker, 2001; Hambleton et al., 1991). Additionally, the test was modified to unify the structure and to reduce the possibility of guessing. In the new structure, the number of items was three correct items out of 12 options for each question. With the new structure, the number of correct items in each question was 25% following a standard four-alternative MCQ test where three alternatives are wrong and one is correct (Roediger & Marsh, 2005). Only 10 original items were accepted for further use and revisions to the test were made (Figure 1, Table 1; Figure S3).

Step 3: Content validity evaluation

Content validity evaluation was conducted after the pilot study to evaluate the item relevance and clarity (DeVon et al., 2007; Figure 1). A two-round expert panel (Lynn, 1986) consisted of the same informants who took part in the focus group interviews described in Step 1 (Vierula, Hupli, et al., 2020). Six experts who had expertise in nursing student selection and reasoning skills were invited, resulting in five experts in the first round and four in the second one. This was considered an acceptable number; an advised minimum for expert panel evaluation is three experts (Lynn, 1986).

First, the panel (*N* = 5) evaluated the test for question and item relevance and clarity using a dichotomous scale (relevant/not relevant, clear/not clear; DeVon et al., 2007; Polit & Beck, 2006). The experts were asked to give rationales and suggest revisions where necessary. The content validity index (CVI) per item (I-CVI; Polit & Beck, 2006) was calculated with percentages following 80% acceptance limit (adequate; Imle & Atwood, 1988). Two questions and three items were modified and continued to the second round together with 19 wrong items with less than 80% consensus. Second, the experts evaluated the clarity of the modified questions and items. The experts also evaluated the 19 wrong items and their relevance compared with the correct items using a dichotomous scale to avoid ambiguity of the correct answers. One response did not follow the instructions, and only three out of four expert evaluations were used. All three experts agreed on the clarity and relevance of the evaluated questions/items, resulting in 100% consensus so further revisions were not needed. The content validity was considered sufficient (Figure 1).

3.4 | Psychometric testing of the ReSki test

3.4.1 | Instrument

As a result of the development process, the ReSki test (version 2; Table 1; Figure S3) was ready for psychometric testing.

3.4.2 | Participants and data collection

Data were collected (28 May 2019) from undergraduate (Bachelor-level) nursing applicants (*N* = 1906, population 18,020) who were taking an electronic entrance examination in six Finnish UASs that were participant HEIs of the ReSSNE project (Haavisto et al., 2019). Applicants received an invitation to participate in the study with the entrance examination invitation before the exam day. Informed consent was obtained from the participants electronically before they started the entrance examination. Only the data collected from the applicants who performed the ReSki test and gave their consent (*N* = 1056) were used for study purposes. The ReSki test (version 2) was administered as part of the learning skills (including language, mathematical- and reasoning skills) domain of the entrance examination.

3.4.3 | Data analysis

The psychometric testing of the ReSki test (version 2) was based on descriptive statistics (means, SDs and ranges), correlation coefficients for subtotal/total scores and IRT modelling with discrimination, difficulty and pseudoguessing parameters. The data were analysed using Statistical Analysis Software (SAS 9.4®; SAS Institute Inc, 2015) and Mplus 8.1. (Muthén & Muthén, 1998–2012). First, frequencies, percentages and central tendency scores (mean, SD, and range) were calculated for participant demographics and descriptive results. Second, a Pearson correlation coefficient was computed to assess the relationships among subtotals (the three questions), between the ReSki test total scores and the domain of learning skills (scores) and between the ReSki test total scores and total entrance examination scores. Third, an IRT analysis using a 2PL model was performed for the analysis of discrimination and difficulty estimates. Twelve items (five in question 1, six in question 2 and one in question three) were removed before the IRT analysis for being extremely easy (>95% of the respondents got the item correct).

After the 2PL analysis, the binary data were further analysed using TestGardener (Li et al., 2019) software (online version) to obtain more detailed ICCs and the third parameter: pseudoguessing. Accumulated evidence should be gathered to develop a valid test (AERA, APA, & NCME, 2014). The TestGardener software uses a new version of the IRT suggested by Ramsay and Wiberg (2017) in which the test performance is presented over non-negative closed intervals (0,100 or 0,*N*). They have established that optimal scoring of binary test data produces improvements in point-wise root-mean-squared error and bias over sum scoring. Further improvement to measure score performance has been demonstrated by using optimal scoring of the full information in option choices. Furthermore, a new algorithm to estimate option response functions and optimal scoring of each examinee's data has been developed (Ramsay et al., 2019). In TestGardener ICCs, the *Y*-axis is the probability of a correct response, whereas the *X*-axis is the ability that is denoted by percent ranks (Figure 2). Because the distribution of test scores varies from one test to another, TestGardener uses percent ranks to make the data more comparable (Ramsay et al., 2020). Thus, the *X*-axis is scaled to only have positive values.

3.4.4 | Ethical considerations

The study followed responsible conduct of research (The Finnish Advisory Board on Research Integrity, 2012). The ethics committee approval was obtained from the ethics committee of the HEI, permission to conduct the study from the participating UASs and the informed consent from the participants. The participants were informed about their anonymity, ability to withdraw and the voluntary nature of the study. Selection results regarding either individual participants or UAS involved in the study were not reported, decreasing the potential risks of exposing the study participants.

4 | RESULTS

4.1 | Participant demographics

Altogether, 1056 undergraduate nursing applicants took the ReSki test and gave their consent to participate in the study. The respondents represented a typical Finnish nursing applicant population: most of the applicants were female and rather young, and the majority (59.5%) were first-time applicants (Table 2).

TABLE 2 Demographic information of participants ($N = 1056^a$)

Variable	N	%	Range	Mean (SD)
Age in years	1050		18–55	24.56 (7.22)
Gender				
Female	904	86.0		
Male	147	14.0		
Background education				
High school	568	54.0		
Vocational school	484	46.0		
Previous degree in higher education				
Yes	93	8.9		
No	953	91.1		
Previous applications for nursing education				
Yes	426	40.5		
No	625	59.5		

[a] Missing values: Age in years (*N* = 6), gender (*N* = 5), background education (*N* = 4), previous degree in higher education (*N* = 10), previous applications for nursing education (*N* = 5).

TABLE 3 Means, SDs and ranges for correct item, subtotal and total scores (*N* = 1056)

Question	Mean	SD	Range
Question 1 Collecting information[a]			
Item 2	0.35	0.23	0–0.5
Item 4	0.45	0.15	0–0.5
Item 7	0.22	0.25	0–0.5
Subtotal	1.02	0.37	0–1.5
Question 2 Processing information[b]			
Item 5	0.32	0.24	0–0.5
Item 6	0.38	0.21	0–0.5
Item 9	0.40	0.20	0–0.5
Subtotal	1.10	0.37	0–1.5
Question 3 Identifying the problem and establishing goals[c]			
Item 6	0.04	0.14	0–0.5
Item 8	0.41	0.19	0–0.5
Item 12	0.15	0.23	0–0.5
Subtotal	0.60	0.35	0–1.5
Total	2.72	0.80	0–4.5

[a] Correct answer = 0.5pt, subtotal scores = 1.5pt. Wrong answer/don't know = 0pt.

[b] Correct answer = 0.5pt, subtotal scores = 1.5pt. Wrong answer/don't know = 0pt.

[c] Correct answer = 0.5pt, subtotal scores = 1.5pt. Wrong answer/don't know = 0pt.

4.2 | Descriptive results and correlations

The lowest mean scores were in question three indicating that it was the most difficult part of the test, whereas questions one and two were easier for the test-takers (Table 3). A mean value close to the centre of the range is considered desirable (DeVellis, 2017). In this study, the mean values of questions 1–3 varied from the centre of the range, but the mean of the total scores was only slightly higher than the centre of the range supporting an appropriate overall difficulty. The SD was similar in all the questions and indicated the variance among the test-takers.

There was a positive and statistically significant correlation among all the variables studied (Table 4). According to the subtotal correlations, if applicants succeeded in one question, they succeeded well in all the test questions, supporting the theoretical basis of the test. According to the total correlations, if applicants succeeded in reasoning skills, they succeeded well in the domain of learning skills and in the overall entrance examination, supporting the assumption that the ReSki test successfully measured cognitive skills.

4.3 | Test discrimination, difficulty and pseudoguessing

First, the 2PL model showed the discrimination and difficulty estimates for each item in all the questions. The discrimination estimates were classified from very high to very low and 0 if not discriminating at all (Baker, 2001; Table 5). Six items were highly discriminative, five were moderate and most of the items (13/24) had a low/very low discrimination estimate. The difficulty estimates were classified from very hard to very easy (Baker, 2001; Hambleton et al., 1991). The results showed that the items ranged from very easy to very hard. Most of the items were easy or even very easy for the test-takers, and the only difficult items were in question three.

TABLE 4 Pearson correlation coefficients (*r*) for subtotal and total scores

Subtotal correlations	*r*	*N*	*p*-value
Question 1[a] scores and question 2[b] scores	0.32	1056	<.0001
Question 1[a] scores and question 3[c] scores	0.27	1056	<.0001
Question 2[b] scores and question 3[c] scores	0.29	1056	<.0001
ReSki test total scores and question 1[a] scores	0.74	1056	<.0001
ReSki test total scores and question 2[b] scores	0.74	1056	<.0001
ReSki test total scores and question 3[c] scores	0.70	1056	<.0001
Total correlations			
ReSki test total scores and learning skills scores (math and language skills)	0.44	1056	<.0001
ReSki test total scores and total entrance examination scores	0.37	1055	<.0001

[a] Collecting information.

[b] Processing information.

[c] Identifying the problem and establishing goals.

Second, the further analysis of the binary data in a new version of IRT (TestGardener software), provided the pseudoguessing parameter and more detailed ICCs for the test items (Figure 2; Table 5; Figures S4–S12). The difficulty estimates were recomputed and classified from easy to difficult according to the 5%, 25%, 50%, 75% and 95% quantiles (Ramsay et al., 2019; Figure 2; Table 5). The results were similar to the difficulty estimates of the 2PL model, indicating that the test was easy for the test-takers, except for question three. Most of the wrong items were easy for the test-takers, and only one wrong item (item 2, question 3) was found to be a functional distractor. The difficulty range of the correct answers varied from easy to difficult.

Third, the pseudoguessing parameter provided information about the opportunity for low-ability applicants to answer items correctly (Ramsay et al., 2019; Tavakol et al., 2014). Theoretically, the pseudoguessing parameter ranges from 0 to 1 but is typically less than 0.30 (Cor et al., 2009; Tavakol et al., 2014). Therefore, the pseudoguessing was considered high if it exceeded the 30% threshold (Figure 2). The pseudoguessing was analysed by observing the cut points in the 5% quantile in the ICCs. This was decided due to the small amount of data available for estimating curve shapes below the bottom and over the top 5% intervals (Ramsay et al., 2019; Figure 2). According to the results, approximately half of the items (13/24) were susceptible for guessing among weaker examinees. However, only one correct item exceeded the 30% threshold indicating that most of the correct items measured the ability of the test-takers rather than the ability to guess. Based on both the ICCs (Figures S4–S12) and the numerical estimates/levels (Table 5), the easy items were more likely wrong items, often demonstrating low discrimination power and high possibility for guessing.

5 | DISCUSSION

This study aimed to develop and test the psychometric properties of the ReSki test for undergraduate nursing student selection purposes. The need to employ valid and reliable admission tools and to construct standardized tests has been identified in higher education (Perkins et al., 2013; Tavakol et al., 2014). Although it has been suggested that reasoning skills be assessed in student selection, the assessment has mainly focused on critical thinking (Vierula, Haavisto, et al., 2020). In this study, an electronical test was developed to assess reasoning as a generic cognitive skill in the selection phase reflecting the requirements of a future profession. To our knowledge, an admission test constructed on the basis of a reasoning process (the test-taker collects and processes the information and finally comes up with a conclusion) has not been developed previously. The psychometrical testing was based

TABLE 5 Item discrimination and difficulty estimates with their standardised errors (SEs) and new version IRT (TestGardener [TG]) item difficulty and pseudoguessing levels[†]

	f[‡]	Discrimination estimates	SE	Difficulty estimates	SE	TG item difficulty levels	TG pseudo-guessing levels
Question 1							
Item 1	685	0.29[e]	0.16	−2.16[5]	1.17	3	>30% threshold
Item 2	733	0.87[c]	0.18	−1.10[4]	0.20	4	<30% threshold
Item 4	953	1.98[a]	0.39	−1.73[4]	0.17	4	>30% threshold
Item 7	463	0.80 [c]	0.27	0.35[3]	0.14	2	<30% threshold
Item 8	933	0.63[d]	0.14	−3.45[5]	0.71	4	>30% threshold
Item 9	984	5.46[a]	1.71	−1.55[4]	0.09	4	>30% threshold
Item 12	780	0.56 [d]	0.13	−1.98[4]	0.45	4	>30% threshold
Question 2							
Item 2	801	0.66[c]	0.27	−1.91[4]	0.69	4	=30% threshold
Item 3	1003	1.32[c]	0.32	−2.76[5]	0.47	4	>30% threshold
Item 5	670	0.51[d]	0.15	−1.15[4]	0.32	3	<30% threshold
Item 6	812	1.23[c]	0.23	−1.25[4]	0.17	4	=30% threshold
Item 9	849	2.32 [a]	0.95	−1.08[4]	0.17	4	<30% threshold
Item 10	671	1.52[b]	0.55	−0.53[4]	0.11	3	<30% threshold
Question 3							
Item 2	515	0.47[d]	0.10	0.09[3]	0.14	2	<30% threshold
Item 3	737	0.14[e]	0.09	−6.18[5]	4.02	4	>30% threshold
Item 4	927	0.38[d]	0.11	−5.42[5]	1.53	4	>30% threshold
Item 5	876	0.26[e]	0.11	−6.13[5]	2.42	4	>30% threshold
Item 6	89	1.44[b]	0.20	2.16[1]	0.20	1	<30% threshold
Item 7	941	0.37[d]	0.11	−5.86[5]	1.73	4	>30% threshold
Item 8	866	0.21[e]	0.13	−7.27[5]	4.51	3	<30% threshold
Item 9	875	0.45[d]	0.12	−3.69[5]	0.90	4	>30% threshold
Item 10	838	0.61[d]	0.12	−2.39[5]	0.43	4	>30% threshold
Item 11	866	0.46[d]	0.11	−3.49[5]	0.83	4	>30% threshold
Item 12	315	20.95[a]	0.65	0.65[2]	0.02	1	<30% threshold

Note: 1 = Difficult (75%–95%), 2 = Moderate to difficult (50%–75%), 3 = Easy to moderate (25%–50%), 4 = Easy (5%–25%).

[a]Very high (>1.70).

[b]High (1.35–1.69).

[c]Moderate (0.65–1.34).

[d]Low (0.35–0.64), [e]Very low (0.1–0.34).

[1]Very hard (>2.0).

[2]Hard (0.5–2.0).

[3]Medium (−0.5 to +0.5).

[4]Easy (−0.5 to −2).

[5]Very easy (below −2).

[†]Correct items highlighted with Bold.

[‡]The frequency refers to the number of applicants (*N* = 1056) who got the item correct.

on IRT modelling instead of the CTT approach to provide in-depth investigation of the item parameters for improving the quality and equality of assessment (Tavakol et al., 2014).

In the test development process, the theoretical basis of the ReSki test was established to ensure sound operationalization of the concept, which has previously been considered challenging (Carbogim et al., 2016; Zuriguel Pérez et al., 2014). The overall development process included structuring two test versions. According to the results, the theoretical basis of the test was supported by the expert evaluations, which indicated acceptable content validity. In the development of test version 2, more categories describing applicants' reasoning skills were identified in the focus group data (Vierula, Hupli, et al., 2020). However, the ReSki test was based on only the most relevant reasoning skills to be assessed in nursing student selection (Vierula, Hupli, et al., 2020). During the development process, we noticed that not all the categories were relevant or even technically possible to include. Developing the test required constant consideration of the possible test difficulty, applicant perspective, features of the electronic platform and using the ReSki test as part of a wider entrance examination (e.g. extent and length of the test). We found that building a test is a complex procedure that differs from a traditional instrument development process (DeVellis, 2017).

The results of the psychometric testing provided support for the reliability and validity of the developed test. Based on the results, the reliability of the ReSki test was supported by item variance (DeVellis, 2017). The subtotal correlations supported the theoretical structure of the test, indicating that decision-making is based on collecting and processing the information. The results supported reasoning skills being cognitive skills as the high-ability applicants performed well both in the ReSki test and in the domain of learning skills. Additionally, the ReSki test identified high-achieving applicants supported by the statistically significant positive total correlations (Schmidt & MacWilliams, 2011). The results indicated the ReSki test as being relatively easy for the applicants. However, question three increased the difficulty level of the test, resulting in the applicant mean performance moving closer to the centre of the range, which supports an acceptable overall test difficulty (DeVellis, 2017). The IRT approach enabled a closer examination of item-level characteristics to specify items that were difficult or easy for the test-takers. Wrong items mostly failed being functional distractors which reflected the poor quality of these items, whereas the quality of the correct items was better. Some wrong items still succeeded in detecting variance among the applicants. For example, one explanation for the difficulty of question three may be that it included a functional distractor.

Overall, the ReSki test is a standardized and thus an objective method to assess nursing applicants reasoning skills. Even though the test included easy items, it also demonstrated the discriminatory value in the study population and can therefore be suggested to be used in the student selection context. A revision of the dysfunctional distractors and their further testing is suggested to achieve more desirable difficulty level of the items. We have learned that developing good items is an iterative process of developing and testing the items (DeVellis, 2017). In the development process, it should also be noted that test-takers may be cleverer than the items (Tavakol et al., 2014). Therefore, sufficient time and resources should be allocated in this iterative process. In the future, the ReSki test could be further developed to other educational contexts, by shifting the focus from generic to competence-specific assessment of reasoning skills.

5.1 | Limitations

Rigorous steps were taken to develop the test, and the development/testing processes were described in detail to provide comprehensive information (DeVon et al., 2007). The limitations of the development process have been stated in earlier publications (Haavisto et al., 2019; Vierula, Haavisto, et al., 2020; Vierula, Hupli, et al., 2020). Moreover, the expert panels used were rather small. Using expert panels may lead to a problem of an inflated estimate of validity when experts endorse most items (Lynn, 1986). However, expert panels were used twice during the development process, including two rounds following a recommendable use of this method (Polit & Beck, 2006). In psychometric testing, the response rate was 55.4%. However, the sample size was big enough for statistical analysis and large in comparison with sample sizes in other similar nursing student selection studies (Vierula, Haavisto, et al., 2020). The sample represented typical characteristics of the population. The IRT analysis was successfully performed, but the discrimination value item 12 (question three) was quite high. Although the results indicate the item being a functional distractor, the results may also indicate the item being too difficult (Tavakol et al., 2014).

6 | CONCLUSION

The ReSki test is a novel, valid objective assessment of undergraduate nursing applicants' reasoning skills. The IRT analysis was successfully conducted, and it provided item-level information that can be used for further development of the test, especially related to the revisions needed for the distractor items and thus the desired adjustment of the difficulty level. The predictive validity and the usability of the test should be considered in future research as well. The study results may benefit HEIs and researchers when developing a test and/or student selection processes. Additionally, the ReSki test could be further developed to other educational contexts shifting its focus from generic reasoning skills to clinical reasoning.

CONFLICT OF INTEREST

No conflict of interest has been declared by the authors.

AUTHOR CONTRIBUTIONS

All authors have participated in commenting and revising the article critically for important intellectual content and have agreed on the final version.

DATA AVAILABILITY STATEMENT

The authors do not wish to share the data.

ORCID

Jonna Vierula https://orcid.org/0000-0002-7538-9837
Kirsi Talman https://orcid.org/0000-0002-2773-9361
Maija Hupli https://orcid.org/0000-0003-1116-410X
Janne Engblom https://orcid.org/0000-0002-5680-0993
Elina Haavisto https://orcid.org/0000-0002-9747-1428

TWITTER

Kirsi Talman @KirsiTalman

REFERENCES

Alfaro-LeFevre, R. (2013). *Critical thinking, clinical reasoning, and clinical judgment: A practical approach* (5th ed.). Elsevier Saunders.

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing* (2014 ed.). American Educational Research Association.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Carbogim, F. C., Oliveira, L. B., & Püschel, V. A. A. (2016). Critical thinking: Concept analysis from the perspective of Rodger's evolutionary method of concept analysis. *Revista Latino-Americana De Enfermagem*, *24*, e2785. https://doi.org/10.1590/1518-8345.1191.2785

Chiu, T.-W., & Camilli, G. (2013). Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement*, *37*(1), 76–86. https://doi.org/10.1177/0146621612459369

Cor, K., Alves, C., & Gierl, M. (2009). Three applications of automated test assembly within a user-friendly modeling environment. *Practical Assessment, Research & Evaluation*, *14*(14), 1–23. https://doi.org/10.7275/02m6-1268

DeVellis, R. F. (2017). *Scale development. Theory and applications* (4th ed.). SAGE Publications.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., Savoy, S., & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, *39*(2), 155–164. https://doi.org/10.1111/j.1547-5069.2007.00161.x

Dimitrov, D., & Shelestak, D. (2003). Psychometric analysis of performance of categories of client needs and nursing process with the NLN diagnostic readiness test. *Journal of Nursing Measurement*, *11*(3), 207–223. https://doi.org/10.1891/jnum.11.3.207.61270

Eurostat. (2019). *Healthcare personnel statistics – Nursing and caring professionals*. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Healthcare_personnel_statistics_-_nursing_and_caring_professionals&oldid=452027#Health_graduates

Facione, N. C., Facione, P. A., & Sanchez, C. A. (1994). Critical thinking disposition as a measure of competent clinical judgment: The Development of the California Critical Thinking Disposition Inventory. *Journal of Nursing Education*, *33*(8), 345–350. https://doi.org/10.3928/0148-4834-19941001-05

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations*. American Philosophical Association.

Francis, R. (2013). *Report of the Mid Staffordshire NHS Foundation Trust public inquiry executive summary*. The Stationary Office. Retrieved from http://webarchive.nationalarchives.gov.uk/20150407084231/http://www.midstaffspublicinquiry.com/report

Graham, C. (2010). Hearing the voices of general staff: A Delphi study of the contributions of general staff to student outcomes. *Journal of Higher Education Policy and Management*, *32*(3), 213–223. https://doi.org/10.1080/13600801003743315

Haavisto, E., Hupli, M., Hahtela, N., Heikkilä, A., Huovila, P., Moisio, E.-L., Yli-Koivisto, L., & Talman, K. (2019). Structure and content of a new entrance exam to select undergraduate nursing students. *International Journal of Nursing Education Scholarship*, *16*(1), 1–15. https://doi.org/10.1515/ijnes-2018-0008

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (1st ed.). SAGE Publications.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, *38*(Suppl9), II28–II42. https://doi.org/10.1097/00005650-200009002-00007

Imle, M. A., & Atwood, J. R. (1988). Retaining qualitative validity while gaining quantitative reliability and validity: Development of the Transition to Parenthood Concerns Scale. *Advances in Nursing Science*, *11*(1), 61–75. https://doi.org/10.1097/00012272-198810000-00007

Kajander-Unkuri, S., Salminen, L., Saarikoski, M., Suhonen, R., & Leino-Kilpi, H. (2013). Competence areas of nursing students in Europe. *Nurse Education Today*, *33*(6), 625–632. https://doi.org/10.1016/j.nedt.2013.01.017

Levett-Jones, T., Hoffman, K., Dempsey, J., Yeun-Sim Jeong, S., Noble, D., Norton, C. A., Roche, J., & Hickey, N. (2010). The 'five rights' of clinical reasoning: An educational model to enhance nursing students' ability to identify and manage clinically 'at risk' patients. *Nurse Education Today*, *30*(6), 515–520. https://doi.org/10.1016/j.nedt.2009.10.020

Li, J., Ramsay, J. O., & Wiberg, M. (2019). TestGardener: A program for optimal scoring and graphical analysis. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. D. Molenaar (Eds.), *Quantitative psychology: 83rd annual meeting of the psychometric society* (pp. 87–94). Springer.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017

MacDuff, C., Stephen, A., & Taylor, R. (2016). Decision precision or holistic heuristic? Insights on on-site selection of student nurses and midwives. *Nurse Education in Practice*, *16*(1), 40–46. https://doi.org/10.1016/j.nepr.2015.06.008

Maillard, P., Dimaggio, G., de Roten, Y., Berthoud, L., Despland, J.-N., & Kramer, U. (2017). Metacognition as a predictor of change in the treatment for borderline personality disorder: A preliminary pilot study. *Journal of Psychotherapy Integration*, *27*(4), 445–459. https://doi.org/10.1037/int0000090

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, *37*(8), 919–931. https://doi.org/10.1080/02602938.2011.586991

McNelis, A. M., Wellman, D. S., Krothe, J. S., Hrisomalos, D. D., McElveen, J. L., & South, R. J. (2010). Revision and evaluation of the Indiana University School of Nursing baccalaureate admission process. *Journal of Professional Nursing*, *26*(3), 188–195. https://doi.org/10.1016/j.profnurs.2010.01.003

Merriam Webster Dictionary. (2020). *Reasoning*. Retrieved from https://www.merriam-webster.com/dictionary/reasoning

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.

National Council on Measurement in Education (NCME). (2017). *Glossary of important assessment and measurement terms – High-stakes testing*. Retrieved from https://web.archive.org/web/20170722194028/http://www.ncme.org/ncme/NCME/Resource_Center/Glossary/NCME/Resource_Center/Glossary1.aspx?hkey=4bb87415-44dc-4088-9ed9-e8515326a061#anchorH

Perkins, A., Burton, L., Dray, B., & Elcock, K. (2013). Evaluation of a Multiple-Mini-Interview protocol used as a selection tool for entry to an undergraduate nursing programme. *Nurse Education Today*, *33*(5), 465–469. https://doi.org/10.1016/j.nedt.2012.04.023

Polit, D. F., & Beck, C. T. (2006). The Content Validity Index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Ramsay, J. O., Li, J., & Wiberg, M. (2020). *Better test scores with TestGardener*. Retrieved from http://www.psych.mcgill.ca/misc/fda/downloads/FDAfuns/OptimalScoreBook.pdf

Ramsay, J. O., & Wiberg, M. (2017). A strategy to replace sum scoring. *Journal of Educational and Behavioral Statistics*, *42*(3), 282–307. https://doi.org/10.3102/1076998616680841

Ramsay, J., Wiberg, M., & Li, J. (2019). Full information optimal scoring. *Journal of Educational and Behavioral Statistics*, *45*(3), 297–315. https://doi.org/10.3102/1076998619885636

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

SAS Institute Inc. (2015). *SAS/SHARE® 9.4: User's guide* (2nd. ed.). Retrieved from https://documentation.sas.com/?docsetId=shrref&docsetTarget=titlepage.htm&docsetVersion=9.4&locale=en

Schmidt, B., & MacWilliams, B. (2011). Admission criteria for undergraduate nursing programs, a systematic review. *Nurse Educator*, *36*(4), 171–174. https://doi.org/10.1097/NNE.0b013e31821fdb9d

Shulruf, B., Bagg, W., Begun, M., Hay, M., Lichtwark, I., Turnock, A., Warnecke, E., Wilkinson, T. J., & Poole, P. J. (2018). The efficacy of medical student selection tools in Australia and New Zealand. *The Medical Journal of Australia*, *208*(5), 214–218. https://doi.org/10.5694/mja17.00400

Simmons, B. (2010). Clinical reasoning: Concept analysis. *Journal of Advanced Nursing*, *66*(5), 1151–1158. https://doi.org/10.1111/j.1365-2648.2010.05262.x

StataCorp. (2017). *Stata statistical software: Release 15*. StataCorp LLC.

Stenlund, T., Lyrén, P.-E., & Eklöf, H. (2017). The successful test taker: Exploring test-taking behavior profiles through cluster analysis. *European Journal of Psychology of Education*, *33*, 403–417.

Sulis, I., & Toland, M. D. (2017). Introduction to multilevel item response theory analysis: Descriptive and explanatory models. *Journal of Early Adolescence*, *37*(1), 85–128. https://doi.org/10.1177/0272431616642328

Tavakol, M., Rahimi-Madiseh, M., & Dennick, R. (2014). Postexamination analysis of objective tests using the three-parameter item response theory. *Journal of Nursing Measurement*, *22*(1), 94–105. https://doi.org/10.1891/1061-3749.22.1.94

The Finnish Advisory Board on Research Integrity. (2012). *Responsible conduct of research and procedures for handling allegations of misconduct in Finland*. Retrieved from https://www.tenk.fi/sites/tenk.fi/files/HTK_ohje_2012.pdf

The Organisation for Economic Co-operation and Development (OECD). (2019). *Health at a glance 2019. OECD indicators*. https://doi.org/10.1787/4dd50c09-en

Timer, J. E., & Clauson, M. I. (2011). The use of selective admissions tools to predict students' success in an advanced standing baccalaureate nursing program. *Nurse Education Today*, *31*(6), 601–606. https://doi.org/10.1016/j.nedt.2010.10.015

University Clinical Aptitude test (UCAT). (2020). *2020 UCAT official guide*. University Clinical Aptitude Test for Medicine and Dentistry. Retrieved from https://www.ucat.ac.uk/media/1409/ucat_guide_2020-covid-19-update.pdf

Vierula, J., Haavisto, E., Hupli, M., & Talman, K. (2020). The assessment of learning skills in nursing student selection: A scoping review. *Assessment & Evaluation in Higher Education*, *45*(4), 496–512. https://doi.org/10.1080/02602938.2019.1666970

Vierula, J., Hupli, M., Talman, K., & Haavisto, E. (2020). Identifying reasoning skills for the selection of undergraduate nursing students: A focus group study. *Contemporary Nurse*, *56*(2), 120–131. https://doi.org/10.1080/10376178.2020.1743732

Waltz, C. F., Strickland, O. L., & Lenz, E. R. (2017). *Measurement in nursing and health research* (5th ed.). Springer Publishing Company.

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171–177. https://doi.org/10.3969/j.issn.1002-0829.2014.03.010

Zuriguel Pérez, E., Lluch Canut, M. T., Falcó Pegueroles, A., Puig Llobet, M., Moreno Arroyo, C., & Roldán Merino, J. (2014). Critical thinking in nursing: Scoping review of the literature. *International Journal of Nursing Practice*, *21*(6), 820–830. https://doi.org/10.1111/ijn.12347

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Vierula J, Talman K, Hupli M, Laakkonen E, Engblom J, Haavisto E. Development and psychometric testing of Reasoning Skills test for nursing student selection: An item response theory approach. *J Adv Nurs*. 2021;77:2549–2560. https://doi.org/10.1111/jan.14799