



RESEARCH

Open Access



Test sensitivity in assessing competencies in nursing education

Eveline Wittmann¹ , Ulrike Weyland², Susan Seeber³, Julia Warwas⁴, Aldin Striković^{1*} , Philine Krebs³, Monja Pohley¹ and Larissa Wilczek²

*Correspondence:

aldin.strikovic@tum.de

¹ TUM School of Social Sciences and Technology, Technical University of Munich (TUM), Arcisstraße 21, 80333 Munich, Germany
Full list of author information is available at the end of the article

Abstract

The identification of effects of vocational education and training conditions on competence development in nursing education requires longitudinal studies. An important precondition is the availability of a test of nursing competence which is economical in use, measures a homogeneous construct throughout years of nursing education and across nursing specializations, and can detect increases in the required competence, hence allowing for sensitive testing. This article describes a cross-sectional study that aimed to optimize a computer-based test measuring nursing competence in care for the elderly—the TEMA test—through the selection of items on the basis of measurement error, differential item functioning, and item difficulty. Evidence of the test sensitivity of the optimized TEMA-L instrument is presented for the second and third year of nursing education. The total sample consisted of $n = 133$ German nursing students from clinical and geriatric nursing. The resulting instrument includes two test booklets consisting of 36 ($WLE = 0.72$) and 35 items ($WLE = 0.70$) respectively for the second and third year of training. The cross-sectional data indicate that the test likely has good properties for sensitive testing of nursing competence in a future longitudinal study. Hence, it might be used to study factors contributing to increases in nursing competence in German VET and serve as an example for similar studies in other countries. Limitations of the current study and related subjects of future research are discussed.

Keywords: Competence development, Computer-based testing, Differential item functioning, Test sensitivity, Item response theory, Nursing education, Vocational education and training

Introduction

It is a common notion that vocational education and training (VET) leads to improvements regarding the competencies of apprentices. However, there is only little empirical knowledge available regarding the effects leading to such improvements in dual VET (Deutscher and Winther 2018), where outcomes may be affected by school-based instruction and in-company training, as well as by complex relations between instruction and learning in both venues. An important question in the use of competence tests for longitudinal studies examining such effects is whether the measurement instruments used are instructionally sensitive, meaning that they can detect improvement related to

the quality of instruction (Naumann et al. 2019) in both the theoretical and the practical sphere (Deutscher and Winther 2018). Naumann et al. (2017) consider *test sensitivity*, understood as the overall variation of test scores across time points or groups, to be a prerequisite for identifying instructional sensitivity. This concept, which is the focus of our paper, implies that the test measures growth on a homogeneous construct over time. In contrast, item sensitivity—understood as a relative measure—can be defined as the degree to which the sensitivity of the respective item deviates from overall test sensitivity; it is usually measured through differential item functioning (DIF; Naumann et al. 2019) and should be low in a sensitive test.

There are only few existing domain-specific competence testing measures for VET that are suitable for larger samples (Abele et al. 2021) and allow for longitudinal application (e.g., Deutscher and Winther 2018). This is particularly true of nursing education, which in Germany is mostly conducted in non-academic settings and, while not officially part of the dual system of VET, is also an example of a dualistic non-academic form of VET with school-based instruction on the one hand and practical on-site training in care institutions on the other (Bals and Wittmann 2009; Lehmann et al. 2014). In this field, most of the internationally available measurement instruments for competencies have consisted for many years of either self-reports (Wu et al. 2015; Yanhua and Watson 2011) or clinical evaluations in real-world settings (e.g., objective-structured clinical examinations; see Solà-Pola et al. 2020). There has been a lack of systematically and consistently developed, valid and reliable assessment instruments in clinical practice (Immonen et al. 2019). Whereas examining nursing competence in real-world situations is preferable to self-reporting in terms of validity (Kajander-Unkuri et al. 2016), it is not only inefficient with larger samples but also deficient in terms of standardization and reliability, particularly in the case of repeated long-term testing. This is likely a reason why longitudinal studies are rare (e.g., Fan et al. 2015). One way to address these issues is through computerized testing, which the National Council Licensure Examination (NCLEX) requires for nursing licensure in the United States. To address standardization issues in these admission examinations for vocational nursing practice, Woo and Dragan (2012) carried out item sensitivity analyses for content relevance to subgroups based on DIF analyses. However, we could not find any study of nursing competence in the international and national literature conducted with the purpose of testing this construct across years of nursing education or even preparing for its sensitive and economical longitudinal testing. We aim to lay the foundation for such testing in the study presented in this paper.

To address issues of valid and reliable testing of nursing competencies in larger samples, we developed a computer-based test on nursing competence in care for the elderly using a video-based situational judgment approach. We reported in Kaspar et al. (2016) on the measurement quality of the TEMA test in a calibration study, using empirical evidence from a cross-sectional large-scale assessment with 402 geriatric nursing students at the end of nursing education. The test construction supports its curricular and content validity to test nursing competence across geriatric and clinical nursing. However, we were not able to examine its suitability for testing across years of nursing education. Hence, the TEMA test could be used reliably to determine and compare the results of apprentices in geriatric nursing at the end of VET across the expected capability range

for students but not (yet) to determine progress throughout VET. In addition, the TEMA test comprises 77 items, requiring almost two hours of testing time, which restricts its economical application in combination with other instruments, such as measures of the quality of VET.

With the cross-sectional study presented in this paper, we therefore aim to further optimize this computerized instrument in two ways. This involves, first, enhanced test economics through a reduction in the number of items and, second, the design of an instrument that allows for tracing progress on a homogeneous core construct of nursing competence in care for the elderly. In preparation for a future longitudinal study, we present evidence of the intended test sensitivity of the TEMA test for the second and third year of nursing education and across nursing specializations (clinical and geriatric nursing). Hence, our research questions in this study are whether it is possible (1) to create an economical short form of the TEMA test providing for acceptable reliability, (2) to maximize test sensitivity by reducing the number of items whose relative item sensitivity deviates substantially from overall test sensitivity by applying differential item functioning (see Naumann et al. 2016, 2017), (3) to create a test enabling us to account for increases in achievement according to years of education, specifically to avoid floor effects. We pursue these targets while at the same time maintaining curricular and content validity and being fair across years of nursing education and nursing specializations. The purpose is to create an economic, reliable, and homogeneous test in which item difficulties balance out across the test for these subgroups, that is, preconditions other than increasing overall achievement on the core construct. With the resulting instrument, it should be possible to examine its aptitude for longitudinal analysis of competence development or to establish instructional sensitivity in a future study, for example by linking test results to the quality of VET (Wittmann et al. 2022; see Naumann et al. 2019).

The TEMA test

Against the background of the increasing relevance of care work for the elderly, implying specific foci such as multimorbidity or cognitive decline, we developed the TEMA test in order to evaluate the learning outcomes for nursing students regarding care for the elderly. To achieve this goal, we proposed a conceptual model of geriatric care competencies to guide the selection of a set of care situations and specific nursing behaviors for competence testing and to define a statistical model for estimating proficiency on the basis of test data. The TEMA test refers to competent action and interaction with care recipients and family members.¹ The instrument is intended to acknowledge care as a continuing mutual relationship with the care recipient and to align with the central elements of the care process, including diagnosis, intervention, and reflection (Kaspar et al. 2016).

The test is provided in the form of a video-based situational judgment test. Since competence assessment relies critically on the adequate representation of situations calling for the required behavior, we defined and validated a sampling space of everyday

¹ Other facets of competent nursing, such as interactions with other health professionals or coping with the care provider's own resources and vulnerabilities, are currently specified both theoretically and empirically as separate facets of nursing competence in the follow-up project EKGe (Wittmann et al. 2022).

demands and challenges in care for elderly persons by means of systematic curricular analysis and expert interviews and refined it on the basis of Hundenborn's (2007) concept of care situations. The test environment provides a set of care situations from three institutional fields of practice covering three major incidents of care for the elderly (dementia, chronic diseases, end of life): (1) long-term group care (LTC) for patients with dementia (dementia hostel), (2) outpatient care (OTC) with a focus on chronic diseases and multimorbidity, and (3) institutional palliative care (PAL); they include five hypothetical care recipients with multiple care needs as cases. Within the fields of practice, we developed an overall set of twelve situations referring to care affordances identified as typical on the basis of curriculum analyses and expert interviews, such as wound and pain management, care planning, nutrition counseling, and emergency measures, among others, providing for item prompts. The situations were transformed into short video sequences of about 1 or 2 min each, with the filming monitored by trained nurses to enhance authenticity of the settings and the acting (Kaspar et al. 2016). Curricular and content validation of the test comprised the breadth of nursing education relevant to care for the elderly in Germany, meaning geriatric and clinical nursing, as well as a generalized program curriculum comprising both specializations since 2020 (see Wittmann et al. 2022). Table 1 provides an overview of the institutional fields of practice, major incidents, and situations.

High test proficiency levels should represent respondents' complex cognitive appraisals, which can serve as a basis for nursing activity in real care situations, including interaction and communication. We thus operationalized them with systematic reference to recognition of emotion, communication of empathetic understanding, and control of emotional expression (for details, see Kaspar and Hartig 2015), as well as bioscience knowledge required for competent diagnostics of the care recipients (Abele et al. 2021).

Item formats cover typical care activities (Fichtmüller and Walter 2007), such as the selection of one of several possible appraisals of situations or states (e.g., Which information will you make use of ...?), behaviors and action plans (e.g., How would you respond to ...? How would you proceed ...? How would you prioritize ...?), or evaluations of observed behavior (e.g., How would you evaluate/interpret ...?). During test construction, the video-based situational stimuli were checked by nursing students in two pilot studies for issues such as undue exaggeration, stereotyping, and lack of consistency with the adjoining test items. To provide for standardized scoring, item responses must be given in closed format (multiple choice, true–false, image map, right order). In Table 2, we list the number of situations and items as well as maximum point scores for each of the curriculum- and content-validated activity fields in the TEMA test. Since respondents can achieve up to three points for items in the true–false format, a maximum score of 95 points is possible for the entirety of 77 items (Kaspar et al. 2016).

To estimate the psychometric qualities in the original calibration study, we asked 402 geriatric nursing students from 24 German schools at the end of VET to respond to the computer-based test. Multi-dimensional item response theory (IRT) modeling served as a means of estimating proficiency. The standardized computer-based testing (CBT) measures nursing students' client-directed care competence with acceptable precision ($WLE=0.76$) in an optimized test version using 64 items, and does so across the whole range of observed proficiency levels. Test items from all proposed institutional fields of

Table 1 Overview of the institutional fields of practice, major care incidents, and situations in the TEMA assessment

Field of practice 1: Long-term care Care and assistance for care recipients with dementia	Field of practice 2: Outpatient care Care and assistance for care recipients with chronic illnesses	Field of practice 3: Institutionalized palliative care Care and assistance for care recipients at the end of life
<ol style="list-style-type: none"> 1. Teamwork in handovers, cooperation, care documentation and planning 2. Biography-oriented personal hygiene, decubitus prophylaxis, interacting and communicating with restricted awareness 3. Acting in an emergency situation (shortness of breath), dealing with conflicts 4. Biography-oriented activity with communication restrictions 5. Dealing with measures restricting freedom and with time pressure 	<ol style="list-style-type: none"> 1. Wound management, hygiene in the home environment, dealing with revulsion 2. Food intake in the event of swallowing disorders, action in an emergency situation (aspiration) 3. Participation in geriatric rehabilitation concepts (e.g., Bobath concept), guidance for relatives 4. Nutritional counseling, blood sugar control and insulin administration 	<ol style="list-style-type: none"> 1. Admission interview, relationship building, pain management 2. Ethical decision-making (refusal to eat) 3. Terminal care and grief counseling, working with relatives, caring for the deceased

Table 2 Number of situations and items and maximum point scores according to activity fields/major care incidents

Activity fields	Number of situations	Number of items	Maximum point scores
(1) Long-term group care (LTC) for patients with dementia (dementia hostel)	5	26	28
(2) Outpatient care (OTC) with a focus on chronic diseases	4	27	38
(3) Institutional palliative care (PAL)	3	24	29
Sum	12	77	95

practice substantially contribute to the overall test reliability, supporting its structural validity (Messick 1987, 1995). As must be noted, the test should be expected to be rather demanding, since test subjects in the original calibration study attained only 45% of the maximum test score (Kaspar et al. 2016).

Another recent cross-sectional study carried out by Ries (2020) on a sample of 408 students in clinical nursing supports the conclusion that the test can be meaningfully applied in clinical nursing as well; it possesses even higher reliability ($WLE=0.87$) than in the calibration study, which may indicate that the test works slightly differently in geriatric nursing than it does in clinical nursing. Similarly to the original calibration study, attainment averaged 44% of the maximum score, raising the question of how to apply the test meaningfully to second- or first-year students while avoiding floor effects. The findings therefore underscore the need to determine how items can be selected for longitudinal testing with the TEMA assessment for the breadth of non-academic nursing education as it relates to the elderly.

Methods

We aim to select items, particularly anchor items, to be able to use the TEMA test efficiently for the purpose of a future longitudinal study, while at the same time validating its fit for students in both clinical nursing related to care for the elderly and geriatric nursing. To achieve this goal, we merged two samples from clinical nursing and geriatric nursing respectively, leading to an overall cross-sectional sample of 133 nursing students. Our sampling strategy involved selecting two classes per year of nursing for each of the subsamples, with second- and third-year data collected at the same schools, in order to create comparable data sets for these nursing education subgroups.² The combined sample slightly skews towards geriatric nursing (57.1% vs. 42.9% as opposed to 51.5% vs. 48.5% in federal nursing student population data). As the test had proved in the previous studies to be rather difficult for students at the end of nursing education, we expected the test might be too difficult for first year students and therefore did not include them in the study. Since class sizes are mostly smaller in the third than in the second year due to dropout, roughly 60% of the students in the overall sample were in the middle of their second year when the test was conducted. Our final sample largely matches the sample from the original calibration study, where 83.1% of respondents were female and the age was heterogeneous, varying from 19 to 54, with an average of 29 years, and 29.4% of respondents originated from families in which languages other

² Due to the April 2020 pandemic-related lockdown we could not entirely complete the strategy in clinical nursing and were unable to gather third year data for the second class.

Table 3 Years of nursing education, gender, and age group for the geriatric nursing and the clinical nursing subsamples (percentage)

	Geriatric nursing		Clinical nursing		Overall sample
	2nd year	3rd year	2nd year	3rd year	
Gender					
m	7 (14.6%)	6 (21.4%)	7 (21.2%)	6 (26.1%)	26 (19.7%)
f	41 (85.4%)	22 (78.6%)	26 (78.8%)	17 (73.9%)	106 (80.3%)
Age group (years)					
< 21	6 (12.5%)	4 (14.3%)	20 (60.6%)	10 (43.5%)	40 (30.3%)
21–25	8 (16.7%)	6 (21.4%)	8 (24.2%)	6 (26.1%)	28 (21.2%)
> 25	34 (70.8%)	18 (64.3%)	5 (15.2%)	7 (30.4%)	64 (48.5%)
Language spoken in the family					
Only German	34 (70.8%)	22 (78.6%)	22 (66.7%)	13 (56.5%)	91 (68.9%)
Other	14 (29.2%)	6 (21.4%)	11 (33.3%)	10 (43.5%)	41 (31.1%)
Sum	48 (100.0%)	28 (100.0%)	33 (100.0%)	23 (100.0%)	132^a (100.0%)
Share of overall sample	36.4%	21.2%	25.0%	17.4%	

^a Data on gender, age, and language spoken lacking for one student

than German were spoken (Döring et al. 2016). With regard to the overall nursing student population, male student nurses were somewhat underrepresented (25.1% in the overall nursing student population; BMBF 2021), and the oldest age group (> 25) was, due to its high share in the geriatric nursing subsample, considerably overrepresented (27.8% in the overall nursing student population; Federal Statistical Office of Germany 2021). While gender played no role as an explaining factor in the original calibration study, proficiency slightly increased with age (Döring et al. 2016). This may point to larger issues of heterogeneity, particularly in geriatric nursing, and should be taken into account when interpreting our results. Table 3 provides an overview of the sample. Congruent with previous findings, respondents averaged 42.26 of a maximum of 95 points on the test.

First, we used one-dimensional Rasch modeling and iteratively excluded items with measurement error in mind, particularly those with low item-total correlation, while preventing one-sided item exclusion by analyzing distribution across the fields of practice and the situations in the TEMA test. Second, we carried out differential item functioning (DIF) analysis to ensure that the test items could be used for sensitive testing in the second and third year of nursing education while being fair across the different nursing education specializations, meaning geriatric and clinical nursing education. To assess subgroup invariance in our sample, we refer to common recommendations from the NEPS study for assessing DIF (Pohl and Carstensen 2012). Thus, we consider absolute differences in estimated difficulties greater than 1 logit to be very strong DIF, and absolute differences between 0.6 and 1 to be worthy of attention for further investigation (Pohl and Carstensen 2012). While the overall test should differentiate between years of nursing education, item DIF should be low. Hence, items that showed a strong subgroup difference were discarded. In the final step, we used the results of these analyses for selecting items to avoid floor effects when testing second- and third-year nursing

apprentices, again applying curricular and content validity as well as reliability considerations. Analyses were carried out with *ConQuest 2.0*.

Results

Item reduction through measurement error minimization

Two items had to be excluded since they were constants, meaning that all item answers were false, and therefore no diagnostic value could be obtained. Rasch scaling of the remaining 75 items led to a reliability score of $WLE = 0.75$. In a first step, we selected items iteratively with reliability in mind in an effort to minimize measurement error (see Appendix Table 7). We considered three measures for this purpose: weighted mean square (WMNSQ), t-statistics, and corrected item-total correlation. Applying common rules of thumb, we considered values of $WMNSQ < 1.15$ as indicative of a close item fit, $1.15 \leq WMNSQ < 1.20$ as a small item misfit, and $WMNSQ \geq 1.20$ as a considerable item misfit (Smith et al. 2008; Gnams and Nusser 2019). By conventional standards, we interpreted t-values greater than +2 or less than -2 as less compatible with the model than expected ($p < 0.05$) (Bond et al. 2021). While WMNSQ lay within a range of 0.8 to 1.2 for all items, suggesting an acceptable model fit (Pohl and Carstensen 2012), the t-value was outside the range of -2.0 to 2.0 for only one item, indicating that the observed data were less consistent with the model than expected ($p < 0.05$) and suggesting that the item should be omitted. In addition to the WMNSQ and t-value, we evaluated item-total correlations. According to common rules of thumb for evaluating the correlations of the item score with the total score, values > 0.20 are considered acceptable (Pohl and Carstensen 2012). With curricular and content validity in mind, we excluded items only if their item-total correlation was less than 0.15. Items with an item-total correlation less than 0.15 may indicate a low discrimination between low and high performers; furthermore, the low correlation can be interpreted as a problem of construct validity. This resulted in the exclusion of 18 items. The remaining item pool contained 56 items with acceptable values on all three criteria and an improved reliability of $WLE = 0.78$.

In order to evaluate whether, while increasing reliability and test economics, items were excluded to the detriment of curricular and content validity, we analyzed the content and the distribution of the excluded items with regard to fields of practice and situations. The content analysis shows that the remaining items continue to reflect the fields of practice and the situations. Table 4 also indicates that items are distributed fairly across fields of practice and situations, indicating that the 56 item pool maintains curricular and content validity.

Differential item functioning

DIF analyses serve to ensure that test items show the same order of difficulty for varying subgroups in the overall sample. DIF exists for an item if its difficulty interacts with subgroup membership (Osterlind and Everson 2009), meaning that test subjects with the same ability score vary in the likelihood that they will answer an item depending on subgroup membership, and implying that a test containing the respective item discriminates

Table 4 Item distributions across fields of practice and situations before and after exclusion

Fields of practice (FP) and situations (Sit)	Number of items before exclusion	Number of items excluded	Number of items after exclusion
FP 1: Care and assistance for care recipients with dementia (long-term care)	26	7	19
Sit. 1.1: Teamwork in handovers, cooperation, care documentation and planning	7	1	6
Sit. 1.2: Biography-oriented personal hygiene, decubitus prophylaxis, interacting and communicating with restricted awareness	8	2	6
Sit. 1.3: Acting in an emergency situation (shortness of breath), dealing with conflicts	4	2	2
Sit. 1.4: Biography-oriented activity with communication restrictions	3	2	1
Sit. 1.5: Dealing with measures restricting freedom and with time pressure	4	0	4
FP 2: Care and assistance for care recipients with chronic illnesses (outpatient care)	27	8	19
Sit. 2.1: Wound management, hygiene in the home environment, dealing with revulsion	7	3	4
Sit. 2.2: Food intake in the event of swallowing disorders, action in an emergency situation (aspiration)	5	1	4
Sit. 2.3: Participation in geriatric rehabilitation concepts (e.g., Bobath concept), guidance for relatives	7	3	4
Sit. 2.4: Nutritional counseling, blood sugar control and insulin administration	8	1	7
FP 3: Care and assistance for care recipients at the end of life (institutionalized palliative care)	24	6	18
Sit. 3.1: Admission interview, relationship building, pain management	13	1	12
Sit. 3.2: Ethical decision-making (refusal to eat)	4	2	2
Sit. 3.3: Terminal care and grief counseling, working with relatives, caring for the deceased	7	3	4
Sum:	77	21	56

against at least one of the subgroups. The main intent in our study is to ensure that no such discrimination occurs for second and third year nursing students and across both geriatric and clinical nursing (Koller et al. 2012). It should be emphasized again that the rationale for this is to create a homogeneous test suitable for tracing overall achievement on the core construct of nursing competence over years of nursing education. Item parameters were fixed at 0. With $N = 133$ nursing students, we conducted both local, that is, item-based, and global, meaning across-item, DIF analyses.

While geriatric nursing students scored slightly higher than their counterparts in clinical nursing, global DIF scores indicated no significant DIF between the geriatric nursing and clinical nursing subgroups (see Appendix Table 10). Item-level analyses indicated that twelve items varied significantly between the program subgroups (A05d, A16d, A19d, A21d, A25d, A47d, A49d, A54d, A70o, A77d, A78d, A81d), six of them strongly (A05d, A19d, A47d, A54d, A70o, A78d), with seven of the items being more difficult for geriatric nursing students (A21d, A25d, A49d, A54d, A70o, A77d, A81d) and five more difficult for clinical nursing students (A05d, A16d, A19d, A47d, A78d; see Appendix

Table 11). Curriculum analysis for the six items with strong DIF (A05d, A19d, A47d, A54d, A70o, A78d) indicated that differences could consistently be attributed to differences in comprehensiveness or explicitness within the curriculum, with the exception of item A78d, which requires bioscience knowledge and might therefore be easier to answer for students from clinical nursing (see Friedel and Treagust 2005). While we therefore excluded these items with strong DIF to generate an instrument for tracing increasing overall achievement in nursing in general, they may still be suitable for a test intended to trace longitudinal development regarding effects of specializations. Differences in the curriculum are of particular significance with regard to item A19d, an item concerning biography-oriented activity, which is granted considerably more time in the geriatric nursing specialization.

As was to be expected, third-year students scored higher on the tasks than second-year students, albeit only slightly. However, DIF analyses regarding the year of education did not lead to significant global DIF either (see Appendix Table 8). On the item level, an inhomogeneous picture arose, with some items being solved more frequently by second-year students and some more frequently by third-year students. Three items showed strong DIF, with two of these items being more frequently solved by third-year students (A03d, A77d) and one more frequently by second-year students (A22d) (see Appendix Table 9). Item A22d refers to restricting freedom in a dementia case, an explicit part of the second-year curriculum for both clinical and geriatric nursing and therefore possibly easier recalled in the second year. Differences for item A03d concerning planning measures for food intake could not be sufficiently traced to curriculum but may be related to the amount of practical experience. In contrast, item A77d, which refers to hypoglycemia symptoms, requires bioscience or medical knowledge and may be better recalled close to the final exam. Overall, this supports the assumption that these items with strong DIF may be excluded from an instrument aiming at measuring progress in nursing competence.

To ensure that the new TEMA-L instrument will be fair for the subgroups and that the competencies of the different subgroups can be compared with each other, we excluded nine items with strong DIF from the test as a result of the DIF analysis (Pohl and Carstensen 2012, p. 12). Table 5 displays the distribution of the omitted and the remaining tasks according to the fields of practice and the situation. Items remain fairly distributed across the fields of practice, and our analyses show that they continue to cover all situational content, except for the biography-oriented activity mentioned above (situation 1.4).

Item selection for measurement in the second and third year of nursing education

Our final step in the construction of a sensitive test was the selection of test items out of the remaining 47 items which can be applied for assessment in the second and third year of nursing education. The criterion for the item selection was that the items chosen should have the most diagnostic information value for different points of measurement in a future longitudinal study, creating an efficient test while at the same time

Table 5 Item distribution across fields of practice and situations before DIF analysis and after exclusion

Fields of practice (FP) and situations (Sit)	Number of items before exclusion	Number of items excluded	Number of items after exclusion
FP 1: Care and assistance for care recipients with dementia (long-term care)	19	4	15
Sit. 1.1: Teamwork in handovers, cooperation, care documentation and planning	6	2	4
Sit. 1.2: Biography-oriented personal hygiene, decubitus prophylaxis, interacting and communicating with restricted awareness	6	0	6
Sit. 1.3: Acting in an emergency situation (shortness of breath), dealing with conflicts	2	0	2
Sit. 1.4: Biography-oriented activity with communication restrictions	1	1	0
Sit. 1.5: Dealing with measures restricting freedom and with time pressure	4	1	3
FP 2: Care and assistance for care recipients with chronic illnesses (outpatient care)	19	3	16
Sit. 2.1: Wound management, hygiene in the home environment, dealing with revulsion	4	0	4
Sit. 2.2: Food intake in the event of swallowing disorders, action in an emergency situation (aspiration)	4	0	4
Sit. 2.3: Participation in geriatric rehabilitation concepts (e.g., Bobath concept), guidance for relatives	4	0	4
Sit. 2.4: Nutritional counseling, blood sugar control and insulin administration	7	3	4
FP 3: Care and assistance for care recipients at the end of life (institutionalized palliative care)	18	2	16
Sit. 3.1: Admission interview, relationship building, pain management	12	2	10
Sit. 3.2: Ethical decision making (refusal to eat)	2	0	2
Sit. 3.3: Terminal care and grief counseling, working with relatives, caring for the deceased	4	0	4
Sum:	56	9	47

maintaining a satisfying level of reliability. To this end, items should be selected for two measurement points in the second year and the third year of education, respectively. Hence, our purpose was to avoid floor effects at a first point of measurement in the second year of nursing education and ceiling effects at the second point of measurement in the third year (Rost 2004). Additionally, we intended a fair distribution specifically across the fields of practice in caring for the elderly, but also across situations, in order to maintain curricular and content validity.

To accomplish this, we first identified a set of 30 items of medium difficulty and good fit values as anchor items. Another 11 items were distributed according to the intended points of measurement, leading to an overall 36 and 35 items respectively for each of them. We chose five items that were solved by a higher share of third-year students and with slightly higher levels of difficulty for measurement in the third year of nursing education. Six items, which were generally solved at a high rate, were selected

Table 6 Reliability, item selection, and distribution for the first and second point of measurement

Measurement second year		Measurement third year	
Item	Situation	Item	Situation
Anchor items (n = 30)			
A01d	1.1	A01d	1.1
A74d	1.1	A74d	1.1
A06d	1.2	A06d	1.2
A08d	1.2	A08d	1.2
A71d	1.2	A71d	1.2
A10d	1.2	A10d	1.2
A14d	1.3	A14d	1.3
A16d	1.3	A16d	1.3
A21d	1.5	A21d	1.5
A24d	1.5	A24d	1.5
A34o	2.1	A34o	2.1
A75o	2.1	A75o	2.1
A28o	2.2	A28o	2.2
A30d	2.2	A30d	2.2
A37d	2.3	A37d	2.3
A38o	2.3	A38o	2.3
A40o	2.3	A40o	2.3
A43d	2.4	A43d	2.4
A45o	2.4	A45o	2.4
A46d	2.4	A46d	2.4
A49d	3.1	A49d	3.1
A51d	3.1	A51d	3.1
A56d	3.1	A56d	3.1
A57d	3.1	A57d	3.1
A81d	3.1	A81d	3.1
A59d	3.2	A59d	3.2
A60d	3.2	A60d	3.2
A63o	3.3	A63o	3.3
A68d	3.3	A68d	3.3
A69o	3.3	A69o	3.3
Measurement point specific items			
A02d	1.1	A73d	1.1
A09d	1.2	A11d	1.2
A25d	2.2	A72d	2.1
A44o	2.4		
A50d	3.1	A55d	3.1
A53d	3.1	A65d	3.3
WLE reliability = 0.72		WLE reliability = 0.70	

for measurement in the second year of nursing education. The results are depicted in Table 6, which shows that the items selected remain fairly distributed across fields of practice and situations, except situation 1.4. While biography-related items still remain part of the instrument, this means that biography-oriented activity was entirely excluded

from the final instrument. Again, it must be emphasized that the purpose of our analysis is to generate an instrument which is fair for the different nursing specializations in the tracing of overall achievement. We would, however, strongly suggest integrating items regarding situation 1.4 into test environments for tracing increases of competence in the geriatric nursing specialization and taking this aspect of nursing education into account for future instrument development. The selected items continue to reflect the situational content of the remaining situations.

To ensure that the test booklets were fair at both intended measurement points, in a final step we performed DIF analyses between the second and third year of nursing education and between the two nursing specializations for the test booklets of the two measurement points. The DIF analyses did not result in a significant global DIF for either of the test booklets. Students in the third year of nursing education and geriatric nursing performed better on the tasks than students in the second year of nursing education and clinical nursing in the test booklet for the third year of nursing education, albeit only slightly (Appendix Table 16). In addition, the item-level results show that no item exhibited strong DIF (see Tables 12, 13, 14, 15, 16, 17, 18, 19). Overall, the reliability of WLE remains at a satisfactory level, with $WLE = 0.72$ for measurement in the second and $WLE = 0.70$ for measurement in the third year of nursing education.

Discussion and limitations

The purpose of this paper was to optimize the TEMA assessment of nursing competencies in care for the elderly in preparation for longitudinal testing in the second and third year of nursing education, under the condition that it must be sufficiently reliable and economical with regard to testing time. To this end, the resulting assessment should be a sensitive test with regard to the core construct: it should reflect learning progress with regard to nursing competence over time without inherently preferring one programmatic group over another or varying the construct according to years of experience. Using Rasch scaling, we managed to identify a body of items which contribute to item-total correlation and reliability, are fair across existing nursing education specializations and across years of nursing education, and generally maintain the curricular and content validity, which formed the basis of the construction of the TEMA test. As the new test version results in only 36 and 35 items respectively per point of measurement, it has the benefit of reducing the number of test items by more than 50%, leading to significant gains in testing time in a longitudinal design.

The sampling strategy involved only two classes per group, and a future study should be broader in scope. Overall, the results of this study must also be interpreted with caution due to the heterogeneity of the sample. While the sample somewhat underrepresents males with regard to the general nursing student population, we consider this to be of lesser importance, as gender was not significantly linked with TEMA measurement of nursing competence in our previous large-scale assessment calibration study,

conducted with a much broader sample of third-year nursing students. Since age was a significant factor linked slightly positively with proficiency in the calibration study, the high share of students older than 25 in the second year of the geriatric nursing subgroup may contribute to third-year students performing only slightly better than second-year students. Generally, while third-year students scored better than second year students in the resulting third-year booklet, global DIF between third- as opposed to second-year nursing education subgroups was not significant. As this is likely due to the small sample size, significant differences might be expected in a larger sample. However, the study was carried out as a cross-sectional study, and longitudinal testing will be necessary to exclude the possibility that the DIF found in this study with regard to the years in the program are an artifact of the cohorts and do not represent individual development throughout nursing education. While the advantage of our cross-sectional study is that it is sample-conserving and avoids repetition effects, our intent in a subsequent study using the TEMA-L instrument is to disentangle such effects using an experimental longitudinal design with a larger sample. Furthermore, while the sample was comprised of nursing students from both clinical and geriatric nursing and retains satisfactory reliability for our overall sample, and although the curricular and content validity of the test was examined against the generalized nursing curriculum in place since 2020, empirical evidence for the new cohort that started in 2020 has yet to be collected and should be a subject of future study. Finally, sensitive items may indicate variations in either school-based instruction or practical training and might be used in future studies in a controlled manner to elucidate possible origins of variance.

Conclusions

In our cross-sectional study, we developed a reliable and economical technology-based assessment of nursing competence in care for the elderly that is sensitive to years of nursing education. These findings lead us to assume that the instrument can be validly applied in a future study for longitudinal testing of nursing competence concerning care-related action and interaction with clients, patients, and family in the second and third year of nursing education. Moreover, this instrument can be used to conduct systematic research on factors improving nursing competence in the context of VET, both in the theoretical and in the practical sphere. Since our cross-sectional approach to testing sensitivity avoids problems of repeated testing attached to longitudinal studies, it might also serve as an empirical example on making sense and use of cross-sectional data in preparing longitudinal test designs.

Appendix

See Tables [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#) and [19](#).

Table 7 Item information for one-dimensional Rasch scaling

Item	Estimate	Standard error	WMNSQ	CI	t-value	Discrimination
A01d	- 0.681	0.083	0.94	(0.87, 1.13)	- 0.9	0.42
A02d	- 1.351	0.086	1.00	(0.76, 1.24)	0.0	0.24
A03d	- 1.226	0.085	0.97	(0.78, 1.22)	- 0.2	0.31
A04o	1.505	0.083	1.02	(0.43, 1.57)	0.2	0.06
A05d	- 0.224	0.083	1.04	(0.93, 1.07)	0.9	0.16
A06d	- 0.338	0.083	1.02	(0.91, 1.09)	0.4	0.17
A07d	0.183	0.082	1.07	(0.93, 1.07)	1.9	0.03
A08d	- 0.450	0.083	1.00	(0.90, 1.10)	0.1	0.19
A09d	- 0.982	0.084	0.98	(0.82, 1.18)	- 0.1	0.26
A10d	- 1.860	0.088	1.03	(0.65, 1.35)	0.2	0.15
A11d	- 0.982	0.084	1.01	(0.82, 1.18)	0.1	0.18
A12d	0.181	0.082	1.00	(0.93, 1.07)	0.1	0.17
A14d	- 0.353	0.083	0.94	(0.91, 1.09)	- 1.3	0.36
A15d	- 0.906	0.084	1.01	(0.84, 1.16)	0.1	0.08
A16d	- 0.831	0.084	0.91	(0.85, 1.15)	- 1.1	0.48
A17o	- 0.490	0.072	1.11	(0.83, 1.17)	1.2	0.12
A18d	0.276	0.083	1.15	(0.93, 1.07)	3.8	- 0.19
A19d	- 1.924	0.088	0.96	(0.63, 1.37)	- 0.2	0.35
A20d	2.439	0.089	0.97	(0.51, 1.49)	- 0.0	0.04
A21d	0.124	0.082	0.99	(0.93, 1.07)	- 0.3	0.20
A22d	0.708	0.083	1.00	(0.88, 1.12)	0.0	0.19
A23d	0.736	0.083	1.01	(0.87, 1.13)	0.1	0.15
A24d	- 0.652	0.083	1.00	(0.87, 1.13)	0.1	0.25
A25d	- 0.642	0.083	0.97	(0.88, 1.12)	- 0.4	0.29
A26d	- 0.378	0.083	1.03	(0.91, 1.09)	0.7	0.10
A27d	- 3.082	0.091	0.96	(0.26, 1.74)	0.0	0.30
A28o	1.265	0.078	0.92	(0.71, 1.29)	- 0.5	0.39
A29d	0.430	0.083	1.04	(0.91, 1.09)	0.9	0.18
A30d	- 0.093	0.083	0.98	(0.93, 1.07)	- 0.4	0.25
A31d	- 0.362	0.083	1.02	(0.91, 1.09)	0.4	0.13
A32d	- 0.159	0.083	1.04	(0.93, 1.07)	1.0	0.12
A34o	0.304	0.068	1.03	(0.86, 1.14)	0.5	0.24
A35o	2.193	0.084	1.11	(0.66, 1.34)	0.6	- 0.00
A37d	0.825	0.084	1.01	(0.86, 1.14)	0.2	0.23
A38o	0.487	0.064	1.12	(0.81, 1.19)	1.2	0.30
A39d	- 2.116	0.088	0.94	(0.59, 1.41)	- 0.2	0.19
A40o	0.951	0.074	1.00	(0.72, 1.28)	0.1	0.32
A41d	1.678	0.087	1.05	(0.70, 1.30)	0.4	0.11
A42d	3.814	0.092	1.26	(0.00, 2.06)	0.6	- 0.03
A43d	0.012	0.082	0.98	(0.94, 1.06)	- 0.7	0.32
A44o	- 0.834	0.075	0.96	(0.78, 1.22)	- 0.3	0.41
A45o	0.442	0.068	0.99	(0.84, 1.16)	- 0.1	0.36
A46d	- 0.354	0.083	0.89	(0.91, 1.09)	- 2.6	0.45
A47d	- 2.661	0.090	0.89	(0.42, 1.58)	- 0.3	0.19
A49d	- 0.206	0.083	0.90	(0.92, 1.08)	- 2.7	0.48
A50d	- 1.572	0.087	0.94	(0.71, 1.29)	- 0.4	0.34
A51d	0.388	0.083	1.01	(0.91, 1.09)	0.2	0.24
A52d	- 0.751	0.084	1.02	(0.86, 1.14)	0.4	0.13
A53d	- 1.272	0.085	0.93	(0.77, 1.23)	- 0.6	0.38

Table 7 (continued)

Item	Estimate	Standard error	WMNSQ	CI	t-value	Discrimination
A54d	1.150	0.085	0.94	(0.80, 1.20)	− 0.6	0.39
A55d	0.416	0.083	0.88	(0.91, 1.09)	− 2.7	0.61
A56d	− 0.149	0.083	1.00	(0.93, 1.07)	0.1	0.26
A57d	0.203	0.083	0.93	(0.93, 1.07)	− 2.0	0.39
A59d	− 0.200	0.083	1.01	(0.93, 1.07)	0.3	0.26
A60d	0.781	0.084	0.95	(0.86, 1.14)	− 0.8	0.36
A61d	1.690	0.087	1.03	(0.70, 1.30)	0.2	0.10
A62o	0.938	0.074	1.05	(0.71, 1.29)	0.4	0.14
A63o	0.220	0.072	1.06	(0.85, 1.15)	0.8	0.19
A65d	− 1.021	0.085	0.95	(0.82, 1.18)	− 0.5	0.37
A66d	0.585	0.083	1.06	(0.89, 1.11)	1.1	0.08
A67d	0.213	0.083	1.03	(0.93, 1.07)	0.8	0.11
A68d	0.141	0.083	0.96	(0.93, 1.07)	− 1.2	0.33
A69o	1.578	0.081	1.02	(0.70, 1.30)	0.2	0.20
A70o	0.622	0.073	1.12	(0.82, 1.18)	1.2	0.13
A71d	− 0.488	0.083	1.01	(0.90, 1.10)	0.2	0.18
A72d	0.035	0.083	1.00	(0.94, 1.06)	0.0	0.24
A73d	1.160	0.085	1.00	(0.80, 1.20)	0.1	0.15
A74d	0.190	0.083	1.00	(0.93, 1.07)	0.1	0.25
A75o	0.600 ^a	0.711	1.06	(0.77, 1.23)	0.6	0.34
A76d	0.265	0.083	1.03	(0.93, 1.07)	0.7	0.15
A77d	− 0.272	0.083	0.99	(0.92, 1.08)	− 0.1	0.29
A78d	− 1.277	0.086	0.96	(0.77, 1.23)	− 0.3	0.27
A79d	0.765	0.085	1.01	(0.86, 1.14)	0.2	0.13
A80d	1.254	0.085	0.98	(0.78, 1.22)	− 0.1	0.18
A81d	− 0.606	0.084	0.99	(0.88, 1.12)	− 0.1	0.21

^a The item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 8 Global DIF analysis to identify subgroup invariance between years of nursing education

	Estimate	Standard error
2nd year of nursing education	− 0.092	0.054
3rd year of nursing education	0.092 ^a	
Chi-square test of parameter equality	2.88	
df	1	
Sig Level	0.089	

^a The item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 9 DIF item analysis to identify subgroup invariance between years of nursing education at the item level

Item	Estimate (2nd year)	Estimate (3rd year)	Standard error
A01d	-0.274	0.274 ^a	0.190
A02d	-0.100	0.100 ^a	0.219
A03d	0.525	-0.525^a	0.246
A05d	0.043	-0.043 ^a	0.183
A06d	-0.121	0.121 ^a	0.184
A08d	-0.214	0.214 ^a	0.185
A09d	-0.145	0.145 ^a	0.200
A10d	-0.054	0.054 ^a	0.254
A11d	0.104	-0.104 ^a	0.205
A14d	0.217	-0.217 ^a	0.188
A16d	0.396	-0.396 ^a	0.211
A19d	0.186	-0.186 ^a	0.276
A21d	-0.069	0.069 ^a	0.182
A22d	-0.529	0.529^a	0.204
A23d	0.111	-0.111 ^a	0.191
A24d	-0.107	0.107 ^a	0.191
A25d	0.127	-0.127 ^a	0.194
A27d	0.051	-0.051 ^a	0.439
A28o	-0.268	0.268 ^a	0.278
A29d	-0.371	0.371 ^a	0.189
A30d	-0.266	0.266 ^a	0.185
A34o	0.162	-0.162 ^a	0.109
A37d	0.048	-0.048 ^a	0.194
A38o	-0.039	0.039 ^a	0.104
A39d	-0.173	0.173 ^a	0.288
A40o	0.071	-0.071 ^a	0.132
A43d	-0.056	0.056 ^a	0.182
A44o	0.086	-0.086 ^a	0.162
A45o	0.136	-0.136 ^a	0.109
A46d	0.117	-0.117 ^a	0.188
A47d	-0.314	0.314 ^a	0.370
A49d	-0.107	0.107 ^a	0.185
A50d	-0.239	0.239 ^a	0.239
A51d	-0.217	0.217 ^a	0.186
A53d	-0.171	0.171 ^a	0.219
A54d	0.389	-0.389 ^a	0.206
A55d	0.225	-0.225 ^a	0.187
A56d	0.042	-0.042 ^a	0.186
A57d	0.179	-0.179 ^a	0.185
A59d	-0.238	0.238 ^a	0.184
A60d	-0.143	0.143 ^a	0.197
A63o	-0.217	0.217 ^a	0.119
A65d	0.140	-0.140 ^a	0.213
A68d	0.063	-0.063 ^a	0.183
A70o	0.080	-0.080 ^a	0.128
A71d	0.164	-0.164 ^a	0.191
A72d	0.144	-0.144 ^a	0.184
A73d	0.252	-0.252 ^a	0.206

Table 9 (continued)

Item	Estimate (2nd year)	Estimate (3rd year)	Standard error
A74d	- 0.024	0.024 ^a	0.182
A77d	0.745	- 0.745^a	0.207
A78d	0.343	- 0.343 ^a	0.241
A79d	- 0.135	0.135 ^a	0.208
A80d	- 0.104	0.104 ^a	0.216
A81d	0.006	- 0.006 ^a	0.193
A69o	- 0.331	0.331 ^a	0.296
A75o	- 0.120 ^a	0.120 ^a	

Separation reliability = 0.192; Chi-square test of parameter equality = 71.38, df = 55, Sig. = 0.068

Significant values are printed in bold

^a The item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 10 Global DIF analysis to identify subgroup invariance between specializations

	Estimate	Standard error
Geriatric nursing	0.048	0.054
Clinical nursing	- 0.048 ^a	
Chi-square test of parameter equality	0.82	
df	1	
Sig Level	0.366	

^a The item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 11 DIF item analysis to identify subgroup invariance between specializations at the item level

Item	Estimate (geriatric nursing)	Estimate (clinical nursing)	Standard error
A01d	- 0.219	0.219 ^a	0.189
A02d	0.038	- 0.038 ^a	0.218
A03d	- 0.192	0.192 ^a	0.209
A05d	- 0.907	0.907^a	0.197
A06d	- 0.308	0.308 ^a	0.184
A08d	0.205	- 0.205 ^a	0.185
A09d	- 0.042	0.042 ^a	0.198
A10d	- 0.059	0.059 ^a	0.250
A11d	- 0.363	0.363 ^a	0.200
A14d	0.086	- 0.086 ^a	0.182
A16d	- 0.440	0.440^a	0.196
A19d	- 0.535	0.535^a	0.268
A21d	0.375	- 0.375^a	0.182
A22d	0.024	- 0.024 ^a	0.189
A23d	- 0.329	0.329 ^a	0.198
A24d	- 0.055	0.055 ^a	0.189
A25d	0.395	- 0.395^a	0.194
A27d	- 0.082	0.082 ^a	0.416
A28o	0.054	- 0.054 ^a	0.241
A29d	- 0.031	0.031 ^a	0.183
A30d	- 0.242	0.242 ^a	0.184
A34o	- 0.113	0.113 ^a	0.109
A37d	0.277	- 0.277 ^a	0.192

Table 11 (continued)

Item	Estimate (geriatric nursing)	Estimate (clinical nursing)	Standard error
A38o	0.003	− 0.003 ^a	0.104
A39d	− 0.112	0.112 ^a	0.284
A40o	0.042	− 0.042 ^a	0.132
A43d	0.301	− 0.301 ^a	0.181
A44o	0.103	− 0.103 ^a	0.154
A45o	− 0.065	0.065 ^a	0.108
A46d	− 0.027	0.027 ^a	0.183
A47d	− 1.121	1.121^a	0.543
A49d	0.454	− 0.454^a	0.189
A50d	0.241	− 0.241 ^a	0.244
A51d	0.019	− 0.019 ^a	0.183
A53d	0.133	− 0.133 ^a	0.220
A54d	0.502	− 0.502^a	0.206
A55d	0.035	− 0.035 ^a	0.185
A56d	0.096	− 0.096 ^a	0.183
A57d	− 0.046	0.046 ^a	0.182
A59d	− 0.354	0.354 ^a	0.185
A60d	0.003	− 0.003 ^a	0.195
A63o	0.045	− 0.045 ^a	0.119
A65d	0.133	− 0.133 ^a	0.206
A68d	0.307	− 0.307 ^a	0.182
A70o	0.661	− 0.661^a	0.144
A71d	0.079	− 0.079 ^a	0.185
A72d	0.347	− 0.347 ^a	0.182
A73d	0.207	− 0.207 ^a	0.205
A74d	− 0.106	0.106 ^a	0.181
A77d	0.489	− 0.489^a	0.188
A78d	− 0.491	0.491^a	0.225
A79d	0.117	− 0.117 ^a	0.206
A80d	0.205	− 0.205 ^a	0.212
A81d	0.402	− 0.402^a	0.196
A69o	− 0.358	0.358 ^a	0.287
A75o	0.216 ^a	− 0.216 ^a	

Separation reliability = 0.581; Chi-square test of parameter equality = 129.72, df = 55, Sig. = 0.000

Significant values are printed in bold

^a The item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 12 Global DIF analysis to identify subgroup invariance between years of nursing education for measurement in the second year

	Estimate	Standard error
2nd year of nursing education	− 0.044	0.054
3rd year of nursing education	0.044 ^a	
Chi-square test of parameter equality	0.67	
df	1	
Sig Level	0.413	

^a The item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 13 DIF item analysis to identify subgroup invariance between years of nursing education at the item level for measurement in the second year

Item	Estimate (2nd year of training)	Estimate (3rd year of training)	Standard error
A01d	- 0.225	0.225 ^a	0.189
A74d	0.025	- 0.025 ^a	0.181
A06d	- 0.070	0.070 ^a	0.183
A08d	- 0.163	0.163 ^a	0.184
A71d	0.209	- 0.209 ^a	0.190
A10d	0.001	- 0.001 ^a	0.255
A14d	0.269	- 0.269 ^a	0.187
A16d	0.450	- 0.450^a	0.210
A21d	- 0.019	0.019 ^a	0.181
A24d	- 0.061	0.061 ^a	0.190
A34o	0.206	- 0.206 ^a	0.109
A75o	- 0.066	0.066 ^a	0.113
A28o	- 0.228	0.228 ^a	0.279
A30d	- 0.225	0.225 ^a	0.183
A37d	0.088	- 0.088 ^a	0.193
A38o	0.009	- 0.009 ^a	0.104
A40o	0.122	- 0.122 ^a	0.132
A43d	- 0.016	0.016 ^a	0.180
A45o	0.183	- 0.183 ^a	0.109
A46d	0.156	- 0.156 ^a	0.186
A49d	- 0.067	0.067 ^a	0.183
A51d	- 0.173	0.173 ^a	0.184
A56d	0.085	- 0.085 ^a	0.184
A57d	0.226	- 0.226 ^a	0.184
A81d	0.057	- 0.057 ^a	0.192
A59d	- 0.193	0.193 ^a	0.183
A60d	- 0.094	0.094 ^a	0.196
A63o	- 0.169	0.169 ^a	0.118
A68d	0.110	- 0.110 ^a	0.182
A69o	- 0.249	0.249 ^a	0.282
A02d	- 0.049	0.049 ^a	0.218
A09d	- 0.094	0.094 ^a	0.199
A25d	0.175	- 0.175 ^a	0.193
A44o	0.122	- 0.122 ^a	0.160
A50d	- 0.199	0.199 ^a	0.235
A53d	- 0.134 ^a	0.134 ^a	

Separation reliability = 0.000; Chi-square test of parameter equality = 30.80, df = 35, Sig. = 0.671

Significant values are printed in bold

^a The item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 14 Global DIF analysis to identify subgroup invariance between specializations for measurement in the second year

	Estimate	Standard error
Geriatric nursing	0.007	0.054
Clinical nursing	- 0.007 ^a	
Chi-square test of parameter equality	0.02	
df	1	
Sig. Level	0.897	

^aThe item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 15 DIF item analysis to identify subgroup invariance between specializations at the item level for measurement in the second year

Item	Estimate (geriatric nursing)	Estimate (clinical nursing)	Standard error
A01d	- 0.261	0.261 ^a	0.187
A74d	- 0.145	0.145 ^a	0.180
A06d	- 0.347	0.347 ^a	0.182
A08d	0.164	- 0.164 ^a	0.184
A71d	0.037	- 0.037 ^a	0.184
A10d	- 0.104	0.104 ^a	0.247
A14d	0.045	- 0.045 ^a	0.181
A16d	- 0.481	0.481^a	0.194
A21d	0.335	- 0.335 ^a	0.180
A24d	- 0.098	0.098 ^a	0.188
A34o	- 0.151	0.151 ^a	0.108
A75o	0.178	- 0.178 ^a	0.112
A28o	0.016	- 0.016 ^a	0.240
A30d	- 0.280	0.280 ^a	0.182
A37d	0.236	- 0.236 ^a	0.191
A38o	- 0.033	0.033 ^a	0.103
A40o	0.006	- 0.006 ^a	0.131
A43d	0.260	- 0.260 ^a	0.180
A45o	- 0.100	0.100 ^a	0.108
A46d	- 0.069	0.069 ^a	0.182
A49d	0.411	- 0.411^a	0.187
A51d	- 0.020	0.020 ^a	0.182
A56d	0.057	- 0.057 ^a	0.181
A57d	- 0.085	0.085 ^a	0.181
A81d	0.361	- 0.361 ^a	0.194
A59d	- 0.394	0.394^a	0.183
A60d	- 0.033	0.033 ^a	0.193
A63o	0.007	- 0.007 ^a	0.119
A68d	0.267	- 0.267 ^a	0.181
A69o	- 0.390	0.390 ^a	0.282
A02d	- 0.006	0.006 ^a	0.216
A09d	- 0.084	0.084 ^a	0.197
A25d	0.354	- 0.354 ^a	0.192
A44o	0.058	- 0.058 ^a	0.152
A50d	0.195	- 0.195 ^a	0.241
A53d	0.093 ^a	- 0.093 ^a	

Separation reliability = 0.332; Chi-square test of parameter equality = 50.48, df = 35, Sig. = 0.044

Significant values are printed in bold

^a The item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 16 Global DIF analysis to identify subgroup invariance between years of nursing education for measurement in the third year

	Estimate	Standard error
2nd year of training	- 0.080	0.047
3rd year of training	0.080 ^a	
Chi-square test of parameter equality	2.87	
df	1	
Sig Level	0.090	

^a The item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 17 DIF item analysis to identify subgroup invariance between years of nursing education at the item level for measurement in the third year

Item	Estimate (2nd year)	Estimate (3rd year)	Standard error
A01d	- 0.259	0.259 ^a	0.188
A74d	- 0.013	0.013 ^a	0.181
A06d	- 0.106	0.106 ^a	0.183
A08d	- 0.198	0.198 ^a	0.184
A71d	0.173	- 0.173 ^a	0.189
A10d	- 0.029	0.029 ^a	0.254
A14d	0.233	- 0.233 ^a	0.187
A16d	0.416	- 0.416 ^a	0.210
A21d	- 0.055	0.055 ^a	0.180
A24d	- 0.094	0.094 ^a	0.190
A34o	0.166	- 0.166 ^a	0.108
A75o	- 0.107	0.107 ^a	0.112
A28o	- 0.269	0.269 ^a	0.276
A30d	- 0.262	0.262 ^a	0.183
A37d	0.049	- 0.049 ^a	0.192
A38o	- 0.033	0.033 ^a	0.104
A40o	0.079	- 0.079 ^a	0.131
A43d	- 0.052	0.052 ^a	0.180
A45o	0.141	- 0.141 ^a	0.108
A46d	0.121	- 0.121 ^a	0.186
A49d	- 0.102	0.102 ^a	0.183
A51d	- 0.210	0.210 ^a	0.184
A56d	0.053	- 0.053 ^a	0.184
A57d	0.189	- 0.189 ^a	0.183
A81d	0.022	- 0.022 ^a	0.191
A59d	- 0.227	0.227 ^a	0.183
A60d	- 0.130	0.130 ^a	0.196
A63o	- 0.206	0.206 ^a	0.118
A68d	0.074	- 0.074 ^a	0.182
A69o	- 0.293	0.293 ^a	0.280
A73d	0.263	- 0.263 ^a	0.205
A11d	0.123	- 0.123 ^a	0.205
A72d	0.152	- 0.152 ^a	0.182
A55d	0.235	- 0.235 ^a	0.186
A65d	0.152 ^a	- 0.152 ^a	

Separation reliability = 0.000; Chi-square test of parameter equality = 32.47, df = 34, Sig. = 0.543

Significant values are printed in bold

^a The item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 18 Global DIF analysis to identify subgroup invariance between specializations for measurement in the third year

	Estimate	Standard error
Geriatric nursing	0.020	0.051
Clinical nursing	− 0.020 ^a	
Chi-square test of parameter equality	0.15	
df	1	
Sig Level	0.695	

^aThe item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Table 19 DIF item analysis to identify subgroup invariance between specializations at the item level for measurement in the third year

Item	Estimate (geriatric nursing)	Estimate (clinical nursing)	Standard error
A01d	− 0.246	0.246 ^a	0.187
A74d	− 0.132	0.132 ^a	0.180
A06d	− 0.334	0.334 ^a	0.182
A08d	0.177	− 0.177 ^a	0.183
A71d	0.050	− 0.050 ^a	0.183
A10d	− 0.089	0.089 ^a	0.247
A14d	0.059	− 0.059 ^a	0.181
A16d	− 0.465	0.465^a	0.194
A21d	0.348	− 0.348 ^a	0.180
A24d	− 0.083	0.083 ^a	0.187
A34o	− 0.137	0.137 ^a	0.108
A75o	0.188	− 0.188 ^a	0.111
A28o	0.026	− 0.026 ^a	0.240
A30d	− 0.268	0.268 ^a	0.182
A37d	0.249	− 0.249 ^a	0.191
A38o	− 0.022	0.022 ^a	0.103
A40o	0.016	− 0.016 ^a	0.131
A43d	0.273	− 0.273 ^a	0.179
A45o	− 0.089	0.089 ^a	0.108
A46d	− 0.054	0.054 ^a	0.181
A49d	0.425	− 0.425^a	0.187
A51d	− 0.007	0.007 ^a	0.182
A56d	0.071	− 0.071 ^a	0.181
A57d	− 0.071	0.071 ^a	0.181
A81d	0.375	− 0.375 ^a	0.194
A59d	− 0.379	0.379^a	0.183
A60d	− 0.021	0.021 ^a	0.193
A63o	0.021	− 0.021 ^a	0.118
A68d	0.281	− 0.281 ^a	0.181
A69o	− 0.381	0.381 ^a	0.278
A73d	0.180	− 0.180 ^a	0.203
A11d	− 0.389	0.389 ^a	0.198
A72d	0.319	− 0.319 ^a	0.180
A55d	0.006	− 0.006 ^a	0.184
A65d	0.105 ^a	− 0.105 ^a	

Separation reliability = 0.401; Chi-square test of parameter equality = 53.69, df = 34, Sig. = 0.017

Significant values are printed in bold

^aThe item parameters were fixed to zero as part of the item analysis. One parameter is fixed by ConQuest by default for model identification purposes

Abbreviations

CBT: Computer-based testing; DIF: Differential item functioning; IRT: Item response theory; LTC: Long-term care; OTC: Outpatient care; PAL: Palliative care; VET: Vocational education and training; WMNSQ: Weighted mean square.

Acknowledgements

Not applicable.

Author contributions

All authors contributed substantially to this work. EW, UW, SE and JW designed the research study and contributed to the interpretation of the data. AS, PK, MP, LW and EW conducted the data collection and provided the data analyses. EW wrote the manuscript; supported by all co-authors. All authors read and approved the final manuscript.

Funding

This research has been funded by the German Federal Ministry of Education and Research (21AP006A).

Availability of data and materials

The datasets used and/or analyzed during the current study are available on reasonable request.

Declarations**Competing interests**

The authors declare that they have no conflict of interest.

Author details

¹TUM School of Social Sciences and Technology, Technical University of Munich (TUM), Arcisstraße 21, 80333 Munich, Germany. ²Professur Für Erziehungswissenschaft mit dem Schwerpunkt Berufspädagogik, Westfälische Wilhelms-Universität Münster, Georgskommende 26, 48143 Münster, Germany. ³Professur für Wirtschaftspädagogik und Personalentwicklung, Georg-August-Universität Göttingen, Platz der Göttinger Sieben 5, 37073 Göttingen, Germany. ⁴Professur für Wirtschaftspädagogik, insb. Theorie und Didaktik beruflicher Bildung, Fruwirthstr. 47, Universität Hohenheim, 70595 Stuttgart, Germany.

Received: 26 November 2021 Accepted: 3 April 2022

Published online: 27 April 2022

References

- Abele S, Deutscher VK, Nickolaus R, Rausch A, Seeber S, Sembill D, Seifried J, Walker F, Weyland U, Winther E, Wittmann E, Wuttke E (2021) Technologiebasierte Kompetenzmessung in der beruflichen Bildung—Die Potenziale der ASCOT-Tests aus Sicht der Curriculum-Instruction-Assessment-Triade [The potential of technology-based competence tests of the ASCOT initiative from the perspective of the curriculum-instruction-assessment triad]. In: Beck K, Oser F (eds.) Aktuelle Resultate und Probleme der Berufsbildungsforschung. Festschrift für Susanne Weber. wbv, Bielefeld, pp. 13–42
- Bals T, Wittmann E (2009) Social and Health Care. In: Baethge M, Arends L (eds.) Feasibility Study VET-LSA. A comparative analysis of occupational profiles and VET programmes in 8 European countries. International report. In cooperation with Schelten A, Müller M, Nickolaus R, Geißel B, Breuer K, Hillen S, Winther E, Bals T, Wittmann E, Barke A. SOFI, Göttingen, pp. 85–98
- Bond TG, Yan Z, Heene M (2021) Applying the Rasch model. Fundamental Measurement in the Human Sciences. Routledge, New York
- Bundesministerium für Bildung und Forschung (BMBF) (2021). Berufsbildungsbericht 2021 [Federal Report on Vocational Education 2021]. Bonn. https://www.bmbf.de/SharedDocs/Publikationen/de/bmbf/3/31684_Berufsbildungsbericht_2021.pdf?__blob=publicationFile&3Bv=5. Accessed 09 Mar 2022.
- Deutscher VK, Winther E (2018) Instructional sensitivity in vocational education. *Learn Instr* 53:21–33. <https://doi.org/10.1016/j.learninstruc.2017.07.004>
- Döring O, Wittmann E, Weyland U, Nauwerth A, Hartig J, Kaspar R, Möllers M, Rechenbach S, Simon J, Worofka I, Kraus K (2016) Technologiebasierte Messung von beruflichen Kompetenzen für die Pflege älterer Menschen: berufsfachliche Kompetenzen, allgemeine Kompetenzen und Kontextfaktoren (TEMA). [Technology-based measurement of professional competencies for elder care: domain-specific competencies, generic competencies, and contextual factors (TEMA)]. In: Beck K, Landenberger M, Oser F (eds.) Technologiebasierte Kompetenzmessung in der beruflichen Bildung. Ergebnisse aus der BMBF-Förderinitiative ASCOT. wbv, Bielefeld, pp. 243–264.
- Fan JY, Wang YH, Chao LF, Jane SW, Hsu LL (2015) Performance evaluation of nursing students following competency-based education. *Nurse Educ* 35(1):97–103. <https://doi.org/10.1016/j.nedt.2014.07.002>
- Federal Statistical Office of Germany (2021) Statistik nach der Pflegeberufe-Ausbildungsfinanzierungsverordnung—2020 [Statistics according to the federal executive order on financing nursing education]. 27.07.2021. Wiesbaden. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Berufliche-Bildung/Publikationen/Downloads-Berufliche-Bildung/pflegeberufe-ausbildungsfinanzierung-vo-5212401207005.xlsx;sessionId=599381BA58D34FE45F54C76BCD730E.live732?__blob=publicationFile. Accessed 27 Feb 2022.
- Fichtmüller F, Walter A (2007) Pflegen lernen. V&R Unipress, Göttingen
- Friedel JM, Treagust DF (2005) Learning bioscience in nursing education: perceptions of the intended and the prescribed curriculum. *Learn Health Soc Care* 4(6):203–216. <https://doi.org/10.1111/j.1473-6861.2005.00104.x>

- Gnambs T, Nusser L (2019) The longitudinal measurement of reasoning abilities in students with special educational needs. *Front Psychol*. <https://doi.org/10.3389/fpsyg.2019.00232>
- Hundenborn G (2007) Fallorientierte Didaktik in der Pflege: Grundlagen und Beispiele für Ausbildung und Prüfung [Case-oriented didactics in nursing]. Urban & Fischer, Munich.
- Immonen K, Oikarainen A, Tomietto M, Kääriäinen M, Tuomikoski AM, Kaučič BM, Filej B, Riklikiene O, Vizcaya-Moreno MF, Perez-Cañaveras RM, De Raeve P, Mikkonen K (2019) Assessment of nursing students' competence in clinical practice: a systematic review of reviews. *Int J Nurs Stud*. <https://doi.org/10.1016/j.ijnurstu.2019.103414>
- Kajander-Unkuri S, Leino-Kilpi H, Katajisto J, Meretoja R, Räisänen A, Saarikoski M, Salminen L, Suhonen R (2016) Congruence between graduating nursing students' self-assessments and mentors' assessments of students' nurse competence. *Collegian* 23(3):303–312. <https://doi.org/10.1016/j.colegn.2015.06.002>
- Kaspar R, Hartig J (2015) Emotional competencies in geriatric nursing: empirical evidence from a computer based large scale assessment calibration study. *Adv Health Sci Educ* 21:105–119. <https://doi.org/10.1007/s10459-015-9616-y>
- Kaspar R, Döring R, Wittmann E, Hartig J, Weyland U, Nauerth A, Möllers M, Rechenbach S, Simon J, Worofka I (2016) Competencies in geriatric nursing: empirical evidence from a computer based large scale assessment calibration study. *Vocat Learn* 9(2):185–206. <https://doi.org/10.1007/s12186-015-9147-y>
- Koller J, Alexandrowicz R, Hatzinger R (2012) Das Rasch-Modell in der Praxis. Eine Einführung mit eRm [The Rasch model in practice: An introduction with eRm]. Facultas, Vienna.
- Lehmann Y, Beutner K, Karge K, Ayerle G, Heinrich S, Behrens J, Landenberger M (2014) Bestandsaufnahme der Ausbildung in den Gesundheitsberufen im europäischen Vergleich [Appraisal of educational programs and qualifications of health occupations/professions: A European comparison]. BMBF, Bonn. <http://doku.iab.de/externe/2014/k140709r02.pdf>. Accessed 04 Nov 2021.
- Messick S (1987) Validity. *ETS Res Rep Ser* 1987(2). <https://doi.org/10.1002/j.2330-8516.1987.tb00244.x>
- Messick S (1995) Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 50(9):741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Naumann A, Hochweber J, Klieme E (2016) A psychometric framework for the evaluation of instructional sensitivity. *Educ Assess* 21(2):89–101. <https://doi.org/10.1080/10627197.2016.1167591>
- Naumann A, Hartig J, Hochweber J (2017) Absolute and relative measures of instructional sensitivity. *J Educ Behav Stat* 42(6):678–705. <https://doi.org/10.3102/1076998617703649>
- Naumann A, Rieser S, Musow S, Hochweber J, Hartig J (2019) Sensitivity of test items to teaching quality. *Learn Instr* 60(1):41–53. <https://doi.org/10.1016/j.learninstruc.2018.11.002>
- Osterlind SJ, Everson HT (2009) *Differential item functioning*, 2nd edn. Sage Publications, Thousand Oaks
- Pohl S, Carstensen C (2012) NEPS technical report—Scaling the data of the competence tests (NEPS Working Paper No. 14). Otto-Friedrich-Universität, Nationales Bildungspanel, Bamberg. https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf. Accessed 04 Nov 2021.
- Ries S (2020) "Kompetenzmessung in der Gesundheits- und Krankenpflegeausbildung" in Anlehnung an das Teilprojekt "TEMA" [Measuring competencies in health care and nursing on the basis of the TEMA project]. Dissertation, Philosophisch-Theologische Hochschule Vallendar. https://kidoks.bs-z-bw.de/frontdoor/deliver/index/docId/1902/file/Dissertation_final_Ries_2020.pdf. Accessed 04 Nov 2021.
- Rost J (2004) *Lehrbuch Testtheorie—Testkonstruktion*, 2nd edn. Hans Huber, Bern
- Smith AB, Rush R, Fallowfield LJ, Velikova G, Sharpe M (2008) Rasch fit statistics and sample size considerations for polytomous data. *BMC Med Res Methodol* 8:33. <https://doi.org/10.1186/1471-2288-8-33>
- Solà-Pola M, Morin-Fraile V, Fabrellas-Padrés N, Raurell-Torreda M, Guanter-Peris L, Guix-Comellas E, Pulpón-Segura AM (2020) The usefulness and acceptance of the OSCE in nursing schools. *Nurse Educ Pract* 43:102736. <https://doi.org/10.1016/j.nepr.2020.102736>
- Wittmann E, Weyland U, Warwas J, Seeber S, Schumann, M (2022) Operationalisierung und Förderung von Bewältigungs- und Kooperationskompetenzen in der Pflegeausbildung—Ansätze im Forschungsprojekt EKGe. [Operationalizing and fostering coping and collaboration competencies in nursing education—approaches in the research project EKGe]. In: Weyland U, Reiber K (eds.), *Berufliches Lehren und Lernen im Zeichen von Professionalisierung—Reflexionen und Analysen zentraler Entwicklungen in den Gesundheitsberufen*. (Beiheft der Zeitschrift für Berufs- und Wirtschaftspädagogik). Franz Steiner, Stuttgart, pp. 49–68.
- Woo A, Dragan M (2012) Ensuring validity of NCLEX® with differential item functioning analysis. *J Nurs Regul* 2(4):29–31. [https://doi.org/10.1016/S2155-8256\(15\)30252-0](https://doi.org/10.1016/S2155-8256(15)30252-0)
- Wu XV, Enskär K, Lee CCS, Wang W (2015) A systematic review of clinical assessment for undergraduate nursing students. *Nurse Educ Today* 35(2):347–359. <https://doi.org/10.1016/j.nedt.2014.11.016>
- Yanhua C, Watson R (2011) A review of clinical competence assessment in nursing. *Nurse Educ* 31(8):832–836. <https://doi.org/10.1016/j.nedt.2011.05.003>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.