Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Development of script concordance test for assessment of clinical reasoning in nursing: Lessons learned regarding construct validity

E.V. Habes [a,*], J.E.M. Kolk [a], M.F.M. Van Brunschot [b], A. Bouwes [a]

[a] Institute of Nursing Studies, HU University of Applied Sciences Utrecht, Utrecht, the Netherlands
[b] Vakbekwaamheid in Zicht, Oisterwijk, the Netherlands

ARTICLE INFO

Keywords:
Script concordance test
Nursing
Clinical reasoning
Education
Curriculum

ABSTRACT

Background: The script concordance test (SCT) has been shown to be an effective tool to assess the clinical reasoning skills of nursing students. Various nursing studies have demonstrated the construct validity of this test. However, studies on the barriers that may impede construct validity during the development process are limited.
Objective: This evaluation describes the barriers to the development of SCT for Bachelor's nursing students and the lessons learned regarding construct validity.
Methods: We conducted a descriptive evaluation of the SCT development and a validation process was performed. The evaluation was based on written comments during the assessment ($N = 327$), a Student's Perspective Questionnaire ($N = 100$), and student feedback during three live review sessions ($N = 27$).
Results: Despite consideration of the guidelines during SCT development, we encountered three main barriers that may impede construct validity. We undertook the necessary efforts to recruit an appropriate expert panel. We overestimated the experts' and students' understanding of the SCT methodology. Additionally, four potential causes of invalid item construction were identified. These possible causes were 'questionable intervention, hypothesis, or investigation', 'blurred data in new information', 'regression to the middle', and 'misinterpretation of the midpoint'.
Conclusion: The three lessons learned are as follows: 1) The recruitment of an appropriate expert panel must not be underestimated. Besides clinical expertise, experts need training in SCT methodology, including awareness of possible pitfalls; 2) SCT training is a prerequisite for SCT as an assessment; and 3) student feedback may offer a deeper understanding of potential hidden script errors and causes for misinterpretation of SCT. Further studies are necessary to identify additional causes which may impede the construct validity of SCT in nursing education.

1. Introduction

Assessing the potential of nursing students to become certified nurses is an educational challenge. The assessment must evaluate theoretical knowledge in addition to competencies such as practical skills and clinical reasoning. Critical thinking and reasoning under uncertain conditions is a crucial element of clinical assessment. The script concordance test (SCT) is a relatively new assessment tool used to evaluate this skill [1,2]. An SCT consists of several scripts that describe a realistic practical situation, known as a vignette. Three

* Corresponding author. Institute of Nursing Studies, HU University of Applied Sciences Utrecht, PO Box 12011, 3501 AA Utrecht, the Netherlands.

E-mail address: vera.habes@hu.nl (E.V. Habes).

https://doi.org/10.1016/j.heliyon.2024.e35151
Received 28 November 2023; Received in revised form 17 July 2024; Accepted 23 July 2024
Available online 24 July 2024

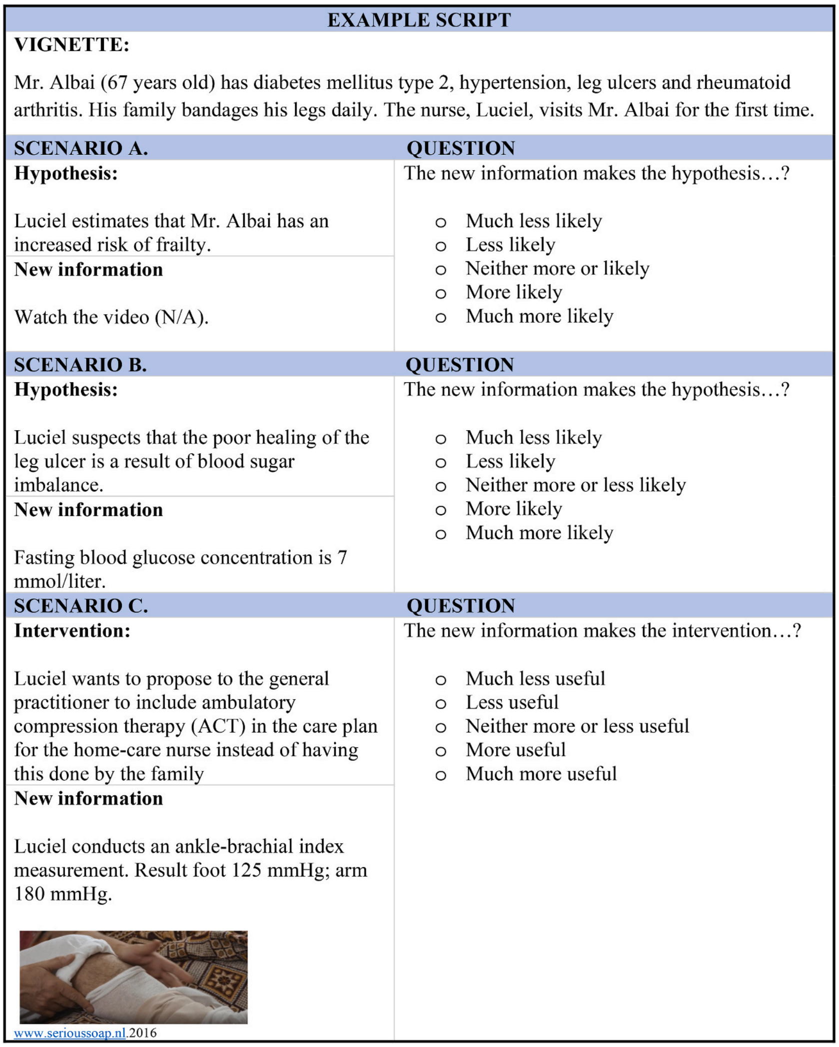
questions are asked regarding a hypothesis, investigation, or intervention, followed by new information (Fig. 1). Students determine how new information modifies a proposed hypothesis (or investigation or intervention) using a 5-point Likert scale [3].

An expert panel validates the correct answers. The expert scoring is based on an aggregate method. This method considers response variability because expert answers vary in situations of uncertainty. The answer provided by the largest number of experts can be regarded as the 'golden standard' of clinical reasoning. Other answers reflect differences in interpretation. This difference is clinically relevant and merits fractional credit [2,4]. Therefore, there is no single 'correct' answer to SCT questions, but some questions yield more credit than others. Fig. 2 illustrates a variety of answers based on Scenarios A and C in Fig. 1. In Scenario A, option 'more likely" receives the most credit (100 %). The student's choice for 'much more likely' receives 75 %. The option 'less likely' gets 50 %. In Scenario C option 'much less useful' receives 100 % credits and option 'less useful' 25 %. Question A yields more credits than question C. The concordance of student and experts scores determines the final result of the assessment.

2. Background

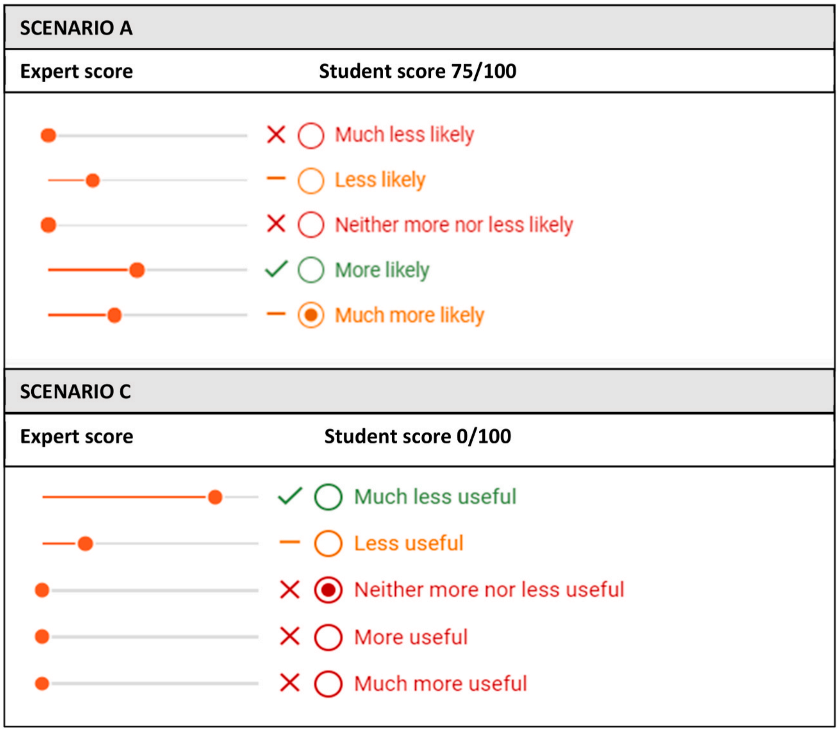
SCT creators designed guidelines for item construction and scoring optimisation to guarantee test validity [5,6]. Various studies have confirmed the reliability and validity of the SCT for nursing students using non-psychometric and psychometric methods [2, 7–12]. Content validity is guaranteed by processing feedback from specialised professionals and content experts [9,10]. Construct validity was demonstrated by increased linear scores between groups with different levels of expertise [6,8). Reliability was statistically demonstrated by an acceptable Cronbach's alpha [2,7,9,10,12]. Item analysis by calculating Rit- and P-values is not used in SCT, possibly because of the variance in credit assignments [13].

We found three medical studies revealing threats to the construct validity and interpretation of SCT scores despite their acceptable



EXAMPLE SCRIPT	
VIGNETTE: Mr. Albai (67 years old) has diabetes mellitus type 2, hypertension, leg ulcers and rheumatoid arthritis. His family bandages his legs daily. The nurse, Luciel, visits Mr. Albai for the first time.	
SCENARIO A.	**QUESTION**
Hypothesis: Luciel estimates that Mr. Albai has an increased risk of frailty.	The new information makes the hypothesis…? ○ Much less likely ○ Less likely ○ Neither more or likely ○ More likely ○ Much more likely
New information Watch the video (N/A).	
SCENARIO B.	**QUESTION**
Hypothesis: Luciel suspects that the poor healing of the leg ulcer is a result of blood sugar imbalance.	The new information makes the hypothesis…? ○ Much less likely ○ Less likely ○ Neither more or less likely ○ More likely ○ Much more likely
New information Fasting blood glucose concentration is 7 mmol/liter.	
SCENARIO C.	**QUESTION**
Intervention: Luciel wants to propose to the general practitioner to include ambulatory compression therapy (ACT) in the care plan for the home-care nurse instead of having this done by the family	The new information makes the intervention…? ○ Much less useful ○ Less useful ○ Neither more or less useful ○ More useful ○ Much more useful
New information Luciel conducts an ankle-brachial index measurement. Result foot 125 mmHg; arm 180 mmHg. www.serioussoap.nl.2016	

Fig. 1. Example Script, 1 vignette and 3 questions.



SCENARIO A

Expert score Student score 75/100

✕ ○ Much less likely

— ○ Less likely

✕ ○ Neither more nor less likely

✓ ○ More likely

— ◉ Much more likely

SCENARIO C

Expert score Student score 0/100

✓ ○ Much less useful

— ○ Less useful

✕ ◉ Neither more nor less useful

✕ ○ More useful

✕ ○ Much more useful

Fig. 2. Example aggregate method, based on scenario A and C of Fig. 1.

psychometric properties [14–16]. Lineberry et al. [15] illustrated how examinees could increase their scores by avoiding extreme scale points, resulting in unjustifiably high marks. In addition, inconsistencies in expert substantiation and discrepancies between substantiation and given scores have been reported [14–16]. These threats to construct validity described in the medical literature have not yet been mentioned in the nursing SCT literature. In addition, possible barriers to nursing SCT development have hardly been described nor discussed [7].

In 2019, nurse educators at the University of Applied Sciences, Utrecht developed an SCT for final-year students. They hypothesised that considering creators' guidelines for development, item construction, and optimisation guarantees valid and reliable SCT. During the development and validation processes, several unexpected barriers were encountered that may have impeded the construct validity.

The aim of our evaluation was to describe the barriers during nursing SCT development, the validation process, and lessons learned regarding construct validity.

3. Method

3.1. Design

A descriptive evaluation of the SCT development and validation process was conducted in three phases of SCT development: construction, implementation, and evaluation (Fig. 3). For the evaluation phase, we used quantitative data from an online Student's Perspective Questionnaire and qualitative data from live review sessions and students' written comments.

3.2. Development phases

3.2.1. Construction phase

A design team comprising eight nursing educators (MSc) was formed. Each pair of designers was specialised in one of four healthcare settings: hospital care, care for older adults, community care, and psychiatric care. The designers were trained by a Ph.D. SCT expert (M. Maas) who recently developed an SCT for physical therapy [17]. The training consisted of a 3-h workshop with an introduction to SCT guidelines [1,6] and feedback on the designers' pilot questions. The distribution of scripts was based on topics related to clinical reasoning as described in the Dutch Body of Knowledge in Nursing [18]. The scripts were evenly distributed across the following four categories: somatic or psychosocial, nursing diagnosis topics, healthcare setting, and age. Each SCT assessment consisted of 20 scripts. Each script was based on a different patient and was labelled according to the four previously described categories. Designers used evolving scripts, incorporating a cumulative story of the three scenarios, thus optimally resembling clinical

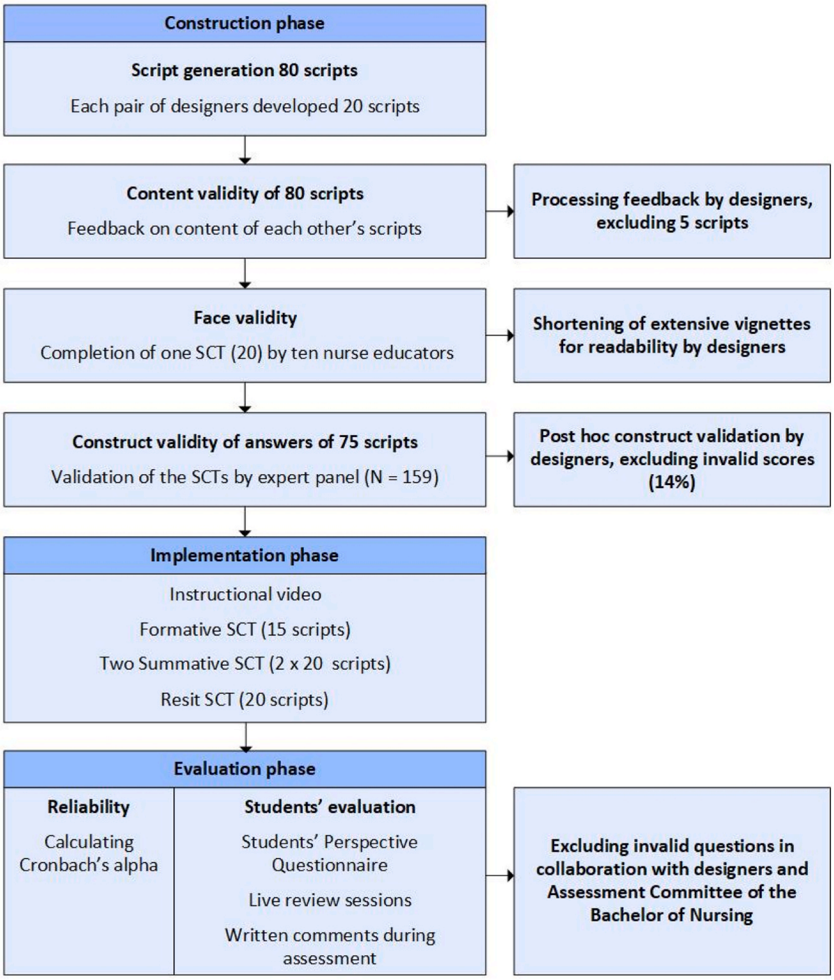

Construction phase
Script generation 80 scripts
Each pair of designers developed 20 scripts
Content validity of 80 scripts
Feedback on content of each other's scripts
Processing feedback by designers, excluding 5 scripts
Face validity
Completion of one SCT (20) by ten nurse educators
Shortening of extensive vignettes for readability by designers
Construct validity of answers of 75 scripts
Validation of the SCTs by expert panel (N = 159)
Post hoc construct validation by designers, excluding invalid scores (14%)
Implementation phase
Instructional video
Formative SCT (15 scripts)
Two Summative SCT (2 x 20 scripts)
Resit SCT (20 scripts)
Evaluation phase
Reliability
Calculating Cronbach's alpha
Students' evaluation
Students' Perspective Questionnaire
Live review sessions
Written comments during assessment
Excluding invalid questions in collaboration with designers and Assessment Committee of the Bachelor of Nursing

Fig. 3. Phases of SCT development

reality [19,20]. The designers aimed to develop four SCTs based on the regulations of the Examination Committee. One SCT was designed as a formative assessment and two as summative assessments, as students were entitled to join a test twice a year. The fourth SCT was designed for a resident.

Four pairs of designers each developed 20 scripts in accordance with their healthcare settings of expertise, for a total of 80 scripts. To ensure the content validity of the scripts, the designers gave feedback on the 60 scripts they had not developed. After processing this feedback, five scripts were excluded because of invalid content, leaving 75 scripts. Next, a face validity test was performed by 10 nursing educators who conducted one SCT (20 scripts) and offered written comments on usability. The main comments of the nurse educators concerned linguistic comments and feedback on the extensive vignettes, which reduced script readability. Designers limited unnecessary information in the vignettes to improve clarity and conciseness.

The designers presented the 75 scripts to an expert panel for construct validation. Experts were asked only to answer questions on familiar topics and substantiate their answers using written 'thought processes' as an alternative to the think-out-loud method [21]. During the validation (1,5 h), experts were offered a 15-min introduction to the SCT methodology, including an instructional video. Nurse educators were invited to participate to meet a minimum of 10 experts per script [22]. For post hoc construct validation, designers analysed experts' scoring and substantiations in pairs and compared experts' substantiations with the latest evidence. Expert substantiation scores deviating from the most recent evidence were also excluded. The number of experts, answers per question, and scores were evaluated.

The researchers consulted the Bachelor of Nursing Assessment Committee to determine scoring criteria. Students earned the highest scores by answering 80 % agreement with experts. The cut-off point of 80 % was based on SCT medical literature showing an expert panel's average score of 80 % [1,23]. A cutoff point of 100 % concordance of the inter-panel score was not eligible because answers needed to reflect the response variability of experts in health care practice.

An information and communication specialist digitised the scripts on the web-based assessment platform iQualify [24]. To prevent incorrect reasoning from one question to the next, students could not return to the previous question but only to the vignette.

3.2.2. Implementation phase

After a pilot study in the spring of 2020, the SCT was introduced in the year 2020/2021 as a final-year assessment. To prepare for the SCT question type, all students watched a 5-min instructional video about the SCT methodology, including an example script. Students had the opportunity to conduct a formative assessment with 15 scripts, incorporating 45 questions as extra preparation.

3.2.3. Evaluation phase

In our nursing department, student feedback was regularly collected after every assessment for internal evaluation. The data collected from the first SCT assessment in March 2021 revealed unexpected potential threats to construct validation, which was a coincidental finding. This finding was an incentive to evaluate students' data more thoroughly using three available data sources: the students' perspective questionnaire, participant observation of students' reflections on their SCT results, and written student feedback generated directly after completing the test.

3.3. Participants and setting

For the construction phase, an expert panel consisting of registered nurses from various healthcare settings was recruited in collaboration with the Dutch professional nurses' organisation, V&VN. The criterion for participation was recent work experience in one of four healthcare settings: hospital care, care for older adults, community care, or psychiatric care.

For student evaluation, we included all final-year students who had undergone the SCT in March 2021 and July 2021. For the qualitative evaluation, we used a convenience sample. Students who voluntarily participated in one of the three live review sessions and those who generated written feedback were included.

3.4. Measurement method

Expert data were collected using iQualify, a digital assessment software tool [24].

The first method of student data collection was an online Student's Perspective Questionnaire on SCT assessment in Evalytics, a digital evaluation tool for higher education [25]. The questions were assessed using a 5-point Likert Scale (Fig. 4). The results of the final (open) questions were categorised into content, student preparation, and satisfaction.

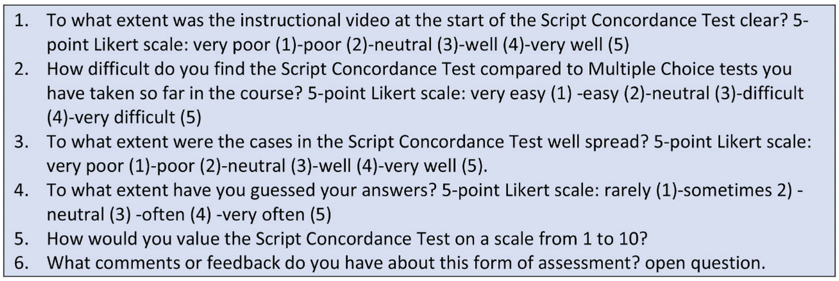
The second method involved participant observation during three live review sessions. These sessions were three weeks after the assessment in the presence of one of the researchers and a clinical reasoning educator. Student participation was voluntary. During these sessions, students reflected on the test results. Students received personal feedback and explanations of the experts' substantiations from educators if needed. After the session, the researcher and educator discussed the problems that impeded students to pass the SCT successfully.

The third method involved students' qualitative feedback immediately after completing the test. Students were able to compare their scores with experts' substantiations and had the opportunity to write comments.

4. Data analysis

The expert scores were calculated using Ku Tools for Excel 2016. The students' baseline characteristics and the reliability of their scores were calculated using SPSS Statistics 29.0 [26]. The reliability was calculated using Cronbach's alpha in the constancy analysis. The results of the Student's Perspective Questionnaire were generated by Evalytics® [25].

Qualitative data from the live review sessions and students' written feedback were analysed by two researchers (E.H. and J.K.) using constant comparative methods [27] and Atlas. ti version 9.0.21 [28]. The researchers independently categorised the data into four themes: negligible comments, linguistic flaws, potential flaws in item construction, and potential flaws in standardisation. Negligible comments were ignored and linguistic flaws were processed directly. Comments on item construction and standardisation were compared with expert substantiation. In the second analysis, the researchers categorised the construction and standardisation flaws into four subthemes: questionable intervention, hypothesis or investigation, blurred data in new information, regression to the middle, and misinterpretation of the midpoint. Consensus was reached through discussions. If an error was discovered, the researchers returned the questions to the designers for clarification and discussion. In case of consensus regarding an error, the question was



1. To what extent was the instructional video at the start of the Script Concordance Test clear? 5-point Likert scale: very poor (1)-poor (2)-neutral (3)-well (4)-very well (5)
2. How difficult do you find the Script Concordance Test compared to Multiple Choice tests you have taken so far in the course? 5-point Likert scale: very easy (1) -easy (2)-neutral (3)-difficult (4)-very difficult (5)
3. To what extent were the cases in the Script Concordance Test well spread? 5-point Likert scale: very poor (1)-poor (2)-neutral (3)-well (4)-very well (5).
4. To what extent have you guessed your answers? 5-point Likert scale: rarely (1)-sometimes 2) -neutral (3) -often (4) -very often (5)
5. How would you value the Script Concordance Test on a scale from 1 to 10?
6. What comments or feedback do you have about this form of assessment? open question.

Fig. 4. Questions Student's Perspective Questionnaire

submitted to the Bachelor of Nursing Assessment Committee. Committee members conducted a final post-analysis of the results. If the committee members agreed upon an invalid construct, the question was removed, and student marks were re-calculated.

5. Results

5.1. Sample size and characteristics

In total, 159 experts joined the panel during the construction phase. An average of 22 experts (SD 5.1) answered each question (range 10–67). A total of 327 final-year students completed the SCT. Their average age was 25.5 years (SD 7.6; median 23). The majority (88 %) is female. A total of 100/327 students (30 %) conducted the Student's Perspective Questionnaire with 76 remarks on the open-ended question. Twenty-seven students (8 %) joined live review sessions. Thirty-nine students (12 %) wrote a total of 125 comments immediately after the assessment.

5.2. Results construction phase

The construct validation of the answers delivered a score of 4986 for 225 questions. All the SCT questions showed response variability and discrimination power, which were required before inclusion in an SCT [28]. During post hoc construct validation, a total of 693 scores (14 %) were excluded. A recommended minimum of 10 experts per script was achieved. Because experts were allowed to skip questions on unfamiliar topics, the answers given could not be traced back to individual experts, resulting in a heterogeneous panel.

5.3. Result evaluation phase

5.3.1. Reliability SCT

The students' calculated average Cronbach's alpha score was 0.6, which is considered average in SCT studies [29].

5.3.2. Student's perspective questionnaire

The results of the questionnaire revealed that students perceived the instructional video as neutral (32 %) or clear (42 %). The students perceived SCT as difficult (50 %) or very difficult (29 %). The topics were considered neutral (41 %) and well-spread (36 %). Students' answers based on guessing were neutral (29 %) or often (35 %) (Fig. 5). Students valued SCT at 7.0 on a scale of 0–10 and appreciated its resemblance to clinical practice. Students mentioned inadequate training in SCT methodology and a lack of SCT education in the curriculum.

5.3.3. Live review sessions

The most important finding of the live review sessions was the students' struggles with the SCT methodology itself. Students made unnecessary mistakes because of misinterpretation of the questions. This finding was confirmed by students' remarks in the questionnaire. The causes of misinterpretation were further clarified by evaluating the students' written comments.

5.3.4. Students' written comments

The evaluation of the students' written comments elicited four potential causes of hidden construct errors. These causes were 'a questionable intervention, hypothesis, or investigation', 'blurred data in new information', 'regression to the middle of scoring', and 'misinterpretation of the midpoint'.

- *Questionable intervention, hypothesis, or investigation:* The first cause of invalid item construction appeared in scripts containing a questionable proposed intervention (hypothesis or investigation). For example, in Scenario C (Fig. 1) the proposed intervention is

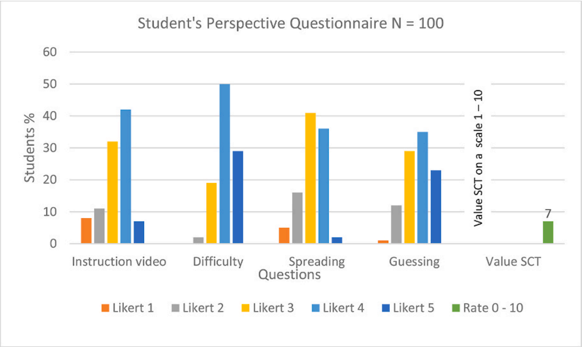


Fig. 5. Student's perspective questionnaire.

ambulatory compression therapy (ACT). This intervention is not intended for discussion. However, some experts and students questioned this intervention as part of their argument for the score, rather than explaining the significance of the new information for the intervention. This misinterpretation is illustrated by Expert 1's substantiation and Student 1's comments, as shown in Fig. 6. Expert 1 presented a case in which ACT was a suitable intervention. Student 1 argued that the ACT was not a suitable intervention. However, both arguments fail to mention the significance of the new information (ankle-brachial index 125/180) for the intervention (ACT).

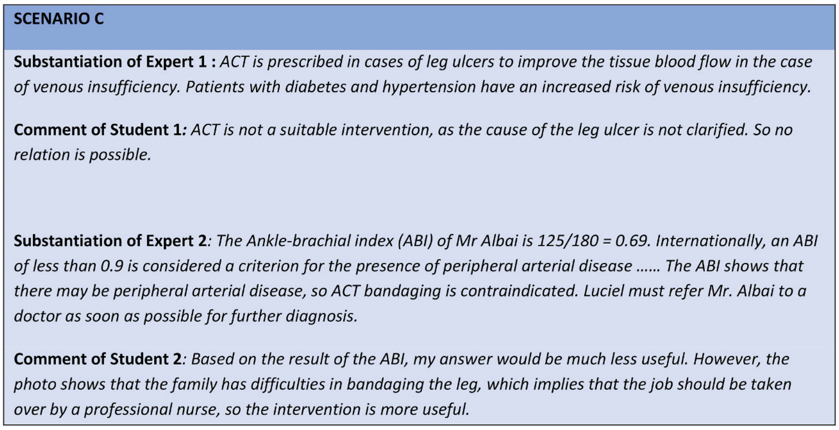
- *Blurred data in new information:* A second cause appears in scripts with new information containing several pieces of new information. For example, the new information in Scenario C in Fig. 1 provides two pieces of information. First, the rate of an ankle-brachial index, and second, the picture with a poorly bandaged leg. This combination of data blurred the reasoning process, as clarified by Student 2's comments, in Fig. 6.
- *Regression to the middle:* experts generally endorsed extreme-scale anchors far less than the midpoint. Hence, expert standardisation showed a regression in the middle (Fig. 7). Extreme scales A (e.g. much less likely) and E (e.g. much more likely) were chosen the least. Although the standardised expert scoring data could have revealed this pattern, we detected this as a potential threat only after reviewing the students' written comments and remarks on the questionnaire. The experts' preference for regression in the middle was hardly explained in their substantiations. When students became aware of this information, mediocre students outperformed students who used the scale as intended by avoiding extreme scores.
- *Misinterpretation of the midpoint:* The experts' unsubstantiated preference for the midpoint led to further investigation of the midpoint's interpretation. The midpoint is described as 'neither more, nor less indicated' [6], as 'it doesn't change your mind' [30], and as 'neutral' (4). Designers adopted the first description in the Dutch SCT. However, students' written comments elicited another interpretation of the midpoint. Students intuitively interpreted the midpoint as 'more *and* less likely', instead of the given 'neither more, nor less likely' description. This misinterpretation led to incorrect answers, as the midpoint meant 'no influence'. In retrospect, this misinterpretation of the midpoint may have caused the experts' unsubstantiated preference for regression to the middle.

6. Discussion

This manuscript provides lessons learned from the evaluation of the SCT development and validation processes. During this process, we identified three main barriers impeding construct validity: underestimation of recruiting an appropriate expert panel, overestimation of SCT understanding by experts and students, and overlooking script errors during the construction phase. We also identified four potential causes of hidden script errors: 'a questionable intervention, hypothesis or investigation', 'blurred data in new information', 'regression to the middle', and 'misinterpretation of the midpoint'.

The underestimation of the effort to recruit an appropriate expert panel can be explained by designers' expectations that joining an expert panel would be perceived as an appealing activity for health professionals, as stated by Lubarsky et al. [4]. Based on the Examination Committee's regulations, the ambition was to develop 80 scripts simultaneously during the construction phase. This appeared to be too ambitious, as it required a large number of experts. Designers' decisions during development to include non-specialised experts and allow them to skip questions led to panel heterogeneity. Dawson [31] described the recruitment of experts as a 'major disadvantage' of SCT, not only because of the quantity but also because of the quality of their expertise. A final cause of underestimation may be our requirement for experts to explain answers using written substantiation. However, this method requires more time and effort than previously indicated. Recruitment of appropriate experts might be easier when experts are asked to validate a limited number of SCT questions focused on their specialisation. Furthermore, the reuse of proven valid and reliable SCT questions and adjustment to the context and country can save time. However, a reanalysis of the adjusted SCT questions remains necessary.

The overestimation of understanding the SCT methodology and possible pitfalls has resulted in the misinterpretation of questions



SCENARIO C

Substantiation of Expert 1 : *ACT is prescribed in cases of leg ulcers to improve the tissue blood flow in the case of venous insufficiency. Patients with diabetes and hypertension have an increased risk of venous insufficiency.*

Comment of Student 1*: ACT is not a suitable intervention, as the cause of the leg ulcer is not clarified. So no relation is possible.*

Substantiation of Expert 2*: The Ankle-brachial index (ABI) of Mr Albai is 125/180 = 0.69. Internationally, an ABI of less than 0.9 is considered a criterion for the presence of peripheral arterial disease ...... The ABI shows that there may be peripheral arterial disease, so ACT bandaging is contraindicated. Luciel must refer Mr. Albai to a doctor as soon as possible for further diagnosis.*

Comment of Student 2*: Based on the result of the ABI, my answer would be much less useful. However, the photo shows that the family has difficulties in bandaging the leg, which implies that the job should be taken over by a professional nurse, so the intervention is more useful.*

Fig. 6. Substantiation of experts and comments of students

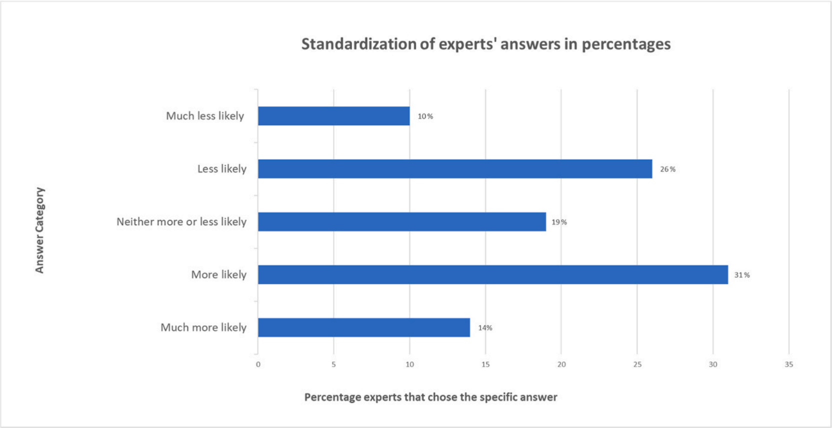


Fig. 7. Standardisation of experts' answers in percentages of 225 questions, after post hoc validation. N = 159.

by experts and students, despite an instructional video. This finding has been confirmed by several studies in the medical field [14–16]. The reported limitations were regression to the middle and inconsistencies between experts' substantiations and scores. In our evaluation, an additional limitation was the misinterpretation of the midpoint by the students and experts. Improved instructions (video), incorporating the possible pitfalls of the methodology, can reduce future misinterpretations of SCT. A more profound solution for students might be to integrate SCT training into the educational nursing curriculum instead of only as a nursing assessment. Deschênes et al. [2] described the theoretical foundation of SCT as a pedagogical strategy that included three elements: SCT as a nursing examination, face-to-face SCT activities, and a digitised educational strategy based on SCT for continuous practice. More SCT activities can offer students extra training in dealing with the uncertainty of clinical nursing practice, as this competence is key to SCT methodology. In particular, the scope of nursing education strives to be holistic, incorporating bio-psychosocial evidence from several sources such as the literature, experts, and patients.

Lubarsky et al. [32] indicated that script-conscious educators are needed, although other educational strategies, such as problem-based learning, think-aloud strategies, and script-based questioning, can also promote script-based clinical reasoning.

The final barrier identified in this evaluation concerns potential hidden script errors in the SCT. Our results highlight these script errors and highlight student feedback as a valuable source for detecting them. By incorporating expert substantiation into the SCT and offering students the possibility of writing comments directly after completing the test, we developed a continuous feedback loop.

The potential initial script errors 'questionable intervention, hypothesis, or investigation' and 'blurred data in new information' were, as far as we know, not mentioned before in the SCT medical literature. One explanation might be that the scope of SCT arguments in nursing is broader than the medical scope, which is mainly based on pathophysiological arguments. In Dutch nursing SCT, experts' substantiations appeared to be based not only on pathophysiological arguments, but also on legal, ethical, psychological, and sociological arguments. This holistic scope of clinical reasoning in nursing may have contributed to the ambiguity of some of the questions. Another explanation may be the evolving script methodology used in the Dutch SCT. The evolving script methodology resembles clinical reality more than the original script methodology [19,20]. The cumulative construction of the three scenarios within one script can differ in terms of the hypothesis, investigation, or intervention. The initial creators of the SCT indicated that scenarios within one script should not differ in terms of hypothesis, investigation, or intervention [6]. The creators warned against the risk that the clinical reasoning process in the first scenario could influence the reasoning process in the next. Although we aimed to prevent this risk by incorporating digital protection, which inhibited the return to the previous scenario, the evolving script methodology may have reinforced the expert panel's divergent focus. The potential initial script errors 'questionable intervention, hypothesis, or investigation' and 'blurred data in new information' could potentially be reduced by avoiding evolving script methodology.

7. Limitations

Some methodological limitations must be mentioned. The first concerns the heterogeneity of the expert panel, which prevents the calculation of construct validity using psychometric analysis. Second, the expert panel's familiarity with the topics was not checked further during the construct validation process. This may have negatively influenced the validity of the experts' answers. To overcome these deficits, the authors are conducting a follow-up study with a sample of scripts from a homogeneous group of specialised experts with strict inclusion criteria.

Second, this evaluation lacked details on the demographics of the participants because the experts' characteristics could not be retrieved. Furthermore, the evaluation results reflected only a few students. The results of the live review sessions were neither recorded nor transcribed. Hence, the analysis of these data was superficial. In addition, the results of the Student's Perspective Questionnaire (30 %) were analysed by only one person. Consequently, we could not draw firm conclusions from our findings because of the subjectivity of interpretation. Finally, the average Cronbach's alpha score was 0.6, which is considered average for SCT studies [29]. One explanation for this could be that the scripts were distributed among different categories.

These limitations affect the validity and generalisability of our conclusions. When designers set stricter inclusion criteria for the expert panel, psychometric analysis could have been performed, and construct errors could have been detected at an earlier phase.

There are also strengths worth mentioning. The reliability of the evaluation was qualitatively enhanced by two researchers reviewing the students' comments. Two researchers independently reviewed the data and compared their findings until a consensus was reached. In addition, the validity of this evaluation was enhanced by triangulation as three sources of data were collected. Furthermore, the standardisation of experts was reanalysed using both experts' substantiations and students' written comments. Students' comments were particularly helpful in detecting the underlying causes of subtle item construction errors and interpreting the midpoint of the Likert scale.

8. Conclusion

We conclude that guaranteeing the construct validity of the SCT in the nursing field might be more difficult than assumed, despite the consideration of guidelines during development. This evaluation taught us three lessons: (1) The recruitment of an appropriate expert panel should not be underestimated. In addition to clinical expertise, experts require training in SCT methods, including awareness of possible pitfalls. (2) SCT training for students is a prerequisite for its use as an assessment tool. (3) A continuous feedback loop based on students' comments can offer a deeper understanding of potential hidden script errors and causes for the misinterpretation of SCT. By describing the barriers to the development of SCT, we hope to share our lessons with other nursing designers of SCT in the future.

Further studies are necessary to identify additional factors which may impede the construct validity of the SCT in nursing education. SCT designers should discuss indisputable criteria for construction and optimisation in the broader context of international nursing education.

Ethical consideration

The Internal Ethical Review Board (Ethical Committee Research ECO) of the University of Applied Sciences Utrecht approved publication of the results in terms of their contribution to the common interest (number 194-000-2022). All the participants cited in this article provided written informed consent.

Data availability statement

The data will be made available upon request. The datasets generated and analysed during this study are not publicly available because the participants, besides those cited in the article, did not provide written consent for their data to be shared publicly.

CRediT authorship contribution statement

E.V. Habes: Writing – review & editing, Writing – original draft. **J.E.M. Kolk:** Writing – review & editing. **M.F.M. van Brunschot:** Software, Formal analysis, Data curation. **A. Bouwes:** Writing – review & editing, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e35151.

References

[1] B. Charlin, R. Gagnon, S. Lubarsky, C. Lambert, S. Meterissian, C. Chalk, J. Goudreau, C. Van der Vleuten, Assessment in the context of uncertainty using the script concordance test: more meaning for scores, Teach. Learn. Med. 22 (3) (2010) 180–186, https://doi.org/10.1080/10401334.2010.488197.

[2] M. Deschênes, D. Létourneau, J. Goudreau, Script concordance approach in nursing education, Nurse Educat. 46 (5) (2021) 103–107, https://doi.org/10.1097/NNE.0000000000001028.

[3] S. Jamieson, Likert scales: how to (ab) use them? Med. Educ. 38 (12) (2004) 1217–1218.

[4] S. Lubarsky, V. Dory, P. Duggan, R. Gagnon, B. Charlin, Script concordance testing: from theory to practice: AMEE guide no. 75, Med. Teach. 35 (3) (2013) 184–193, https://doi.org/10.3109/0142159X.2013.760036.

[5] B. Charlin, R. Gagnon, J. Pelletier, M. Coletti, G. Abi-Rizk, C. Nasr, E. Sauvé, C. Van der Vleuten, Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel, Med. Educ. 40 (9) (2006) 848–854, https://doi.org/10.1111/j.1365-2929.2006.02541.x.

[6] J.P. Fournier, A. Demeester, B. Charlin, Script concordance tests: guidelines for construction, BMC Med. Inf. Decis. Making 8 (18) (2008), https://doi.org/10.1186/1472-6947-8-18.

[7] T. Dawson, L. Comer, M.A. Kossick, J. Neubrander, Can script concordance testing be used in nursing education to accurately assess clinical reasoning skills? J. Nurs. Educ. 53 (5) (2014) 281–286, https://doi.org/10.3928/01484834-20140321-03.

[8] M. Deschênes, B. Charlin, R. Gagnon, J. Goudreau, Use of a script concordance test to assess development of clinical reasoning in nursing students, J. Nurs. Educ. 50 (7) (2011) 381–387, https://doi.org/10.3928/01484834-20110331-03.

[9] G. Kaur, Assessment of clinical reasoning skills among BSc nursing students: script concordance test, Nurs. J. India 112 (4) (2021) 153–156.

[10] P. Srikan, S. Tachvises, Development and psychometric testing of the gerontological nursing clinical reasoning scale, Pac Rim Int J Nurs Res 23 (3) (2019) 243–257.

[11] C. Redmond, A. Jayanth, S. Beresford, L. Carroll, A.N.B. Johnston, Development and validation of a script concordance test to assess biosciences clinical reasoning skills: a cross-sectional study of 1st year undergraduate nursing students, Nurse Educ. Today 119 (2022) 105615, https://doi.org/10.1016/j.nedt.2022.105615.

[12] W. Tapaneeyakorn, J. Kosolchuenvijit, K. Anonrath, S. Wannasuntad, P. Smith, Factors affecting clinical reasoning of nursing students at Boromarajonani College of Nursing Bangkok, J Health Sci Res 10 (2016) 70–77.

[13] H. Van Berkel, A. Bax, D. Joosten-ten Brinke, Toetsen in Het Hoger Onderwijs, Bohn Stafleu & van Loghum, 2017.

[14] N. Gawad, T.J. Wood, L. Cowley, I. Raiche, The cognitive process of test takers when using the script concordance test rating scale, Med. Educ. 54 (4) (2020) 337–347, https://doi.org/10.1111/medu.14056.

[15] M. Lineberry, C.D. Kreiter, G. Bordage, Threats to validity in the use and interpretation of script concordance test scores, Med. Educ. 47 (12) (2013) 1175–1183, https://doi.org/10.1111/medu.12283.

[16] A. Power, J.F. Lemay, S. Cooke, Justify your answer: the role of written think aloud in script concordance testing, Teach. Learn. Med. 29 (1) (2017) 59–67, https://doi.org/10.1080/10401334.2016.1217778.

[17] N.M. Otterman, M. Maas, S.K. Schiemanck, P.J. Van der Wees, G. Kwakkel, Development and validity of an innovative test to assess guideline-consistent clinical reasoning by physical therapists in stroke rehabilitation, J. Rehabil. Med. 51 (6) (2019) 418–425, https://doi.org/10.2340/16501977-2562.

[18] J. Lambregts, A. Grotendorst, C. van Merwijk (Eds.), Bachelor of Nursing 2020: Een Toekomstbestendig Opleidingsprofiel 4.0, Springer, 2016.

[19] S. Cooke, J.F. Lemay, T. Beran, Evolutions in clinical reasoning assessment: the evolving script concordance test, Med. Teach. 39 (8) (2017) 828–835, https://doi.org/10.1080/0142159X.2017.1327706.

[20] H. Molkenboer, Toetsen volgens de toetscyclus, Bureau voor Toetsen & Beoordelen, 2015.

[21] J. Goudreau, L. Boyer, D. Létourneau, Clinical nursing reasoning in nursing practice: a cognitive learning model based on a think aloud methodology, QANE-AFI 1 (1) (2014) 4, https://doi.org/10.17483/2368-6669.1009.

[22] R. Gagnon, B. Charlin, M. Coletti, E. Sauvé, C. Van der Vleuten, Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? Med. Educ. 39 (3) (2005) 284–291. https://doi:10.1111/j.1365-2929.2005.02092.x.

[23] V. Dory, R. Gagnon, D. Vanpee, B. Charlin, How to construct and implement script concordance tests: insights from a systematic review, Med. Educ. 46 (6) (2012) 552–563, https://doi.org/10.1111/j.1365-2923.2011.04211.x.

[24] iQualify, iQualify 6.1.0, [Software], iQualify 6.1.0, 2022. https://www.iqualify.com. (Accessed 6 August 2022).

[25] Evalytics, [Software], 2022. https://evalytics.nl. (Accessed 11 November 2022).

[26] IBM SPSS Statistics for Windows. [Software]. Version 28.0. Armonk, NY: IBM.

[27] B. Glaser, A. Strauss, Discovery of Grounded Theory: Strategies for Qualitative Research, Routledge, 2017.

[28] ATLAS.ti Scientific Software Development [Qualitative data analysis software], Version 9.0.21, for Windows [Software], 2021. https://atlasti.com.

[29] M.F. Deschênes, J. Goudreau, Addressing the development of both knowledge and clinical reasoning in nursing through the perspective of script concordance: an integrative literature review, JNEP 7 (12) (2017) 29–38, https://doi.org/10.5430/jnep.v7n12p28.

[30] R. Gagnon, S. Lubarsky, C. Lambert, B. Charlin, Optimization of answer keys for script concordance testing: should we exclude deviant panelists, deviant responses, or neither? Adv. Health Sci. Educ. 16 (5) (2011) 601–608, https://doi.org/10.1007/s10459-011-9279-2.

[31] T.E. Dawson, Can Script Concordance Testing Be Utilized in Nursing Education to Accurately Assess Clinical Reasoning Skills? Western Carolina University, 2012. Doctoral dissertation.

[32] S. Lubarsky, V. Dory, M. Audétat, E. Custers, B. Charlin, Using script theory to cultivate illness script formation and clinical reasoning in health professions education, Can Med Educ J 6 (2) (2015) e61–e70, https://doi.org/10.36834/cmej.36631.