Validity evidence supporting clinical skills assessment by artificial intelligence compared with trained clinician raters

Johnsson, Vilma; Søndergaard, Morten Bo; Kulasegaram, Kulamakan; Sundberg, Karin; Tiblad, Eleonor; Herling, Lotta; Petersen, Olav Bjørn; Tolsgaard, Martin G.

Published in:
Medical Education

DOI:
10.1111/medu.15190

Publication date:
2024

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Validity evidence supporting clinical skills assessment by artificial intelligence compared with trained clinician raters

Vilma Johnsson1,2 | Morten Bo Søndergaard3,4 | Kulumakan Kulasegaram5,6 | Karin Sundberg1 | Eleonor Tiblad7,8 | Lotta Herling7,9 | Olav Bjørn Petersen1,10 | Martin G. Tolsgaard1,3,10

1Center for Fetal Medicine, Department of Obstetrics, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
2Faculty of Health and Medical Science, University of Copenhagen, Copenhagen, Denmark
3Copenhagen Academy for Medical Education and Simulation, Copenhagen, Denmark
4Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
5Department of Family and Community Medicine and Scientist, Wilson Centre, Toronto, Ontario, Canada
6Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
7Center for Fetal Medicine, Karolinska University Hospital, Stockholm, Sweden
8Clinical Epidemiology Division, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden
9Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Stockholm, Sweden
10Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

Correspondence
Vilma Johnsson, Center for Fetal Medicine, Department of Obstetrics, Copenhagen University Hospital, Rigshospitalet, Blejamsvej 9, 2000 Copenhagen, Denmark. Email: vilmanykvist@gmail.com

Funding information
Vilma Johnsson has received a grant from AMEE and the Danish Ministry of Higher Education and Science. The project was supported by the Novo Nordisk Foundation (grant NNFSA170030576).

Abstract

Background: Artificial intelligence (AI) is becoming increasingly used in medical education, but our understanding of the validity of AI-based assessments (AIBA) as compared with traditional clinical expert-based assessments (EBA) is limited. In this study, the authors aimed to compare and contrast the validity evidence for the assessment of a complex clinical skill based on scores generated from an AI and trained clinical experts, respectively.

Methods: The study was conducted between September 2020 to October 2022. The authors used Kane’s validity framework to prioritise and organise their evidence according to the four inferences: scoring, generalisation, extrapolation and implications. The context of the study was chorionic villus sampling performed within the simulated setting. AIBA and EBA were used to evaluate performances of experts, intermediates and novice based on video recordings. The clinical experts used a scoring instrument developed in a previous international consensus study. The AI used convolutional neural networks for capturing features on video recordings, motion tracking and eye movements to arrive at a final composite score.

Results: A total of 45 individuals participated in the study (22 novices, 12 intermediates and 11 experts). The authors demonstrated validity evidence for scoring, generalisation, extrapolation and implications for both EBA and AIBA. The plausibility of assumptions related to scoring, evidence of reproducibility and relation to different training levels was examined. Issues relating to construct underrepresentation, lack of explainability, and threats to robustness were identified as potential weak links in the AIBA validity argument compared with the EBA validity argument.

Conclusion: There were weak links in the use of AIBA compared with EBA, mainly in their representation of the underlying construct but also regarding their explainability and ability to transfer to other datasets. However, combining AI and clinical expert-based assessments may offer complementary benefits, which is a promising subject for future research.
1 | INTRODUCTION

Assessments are critical for competency-based medical education, particularly in mastery learning of clinical skills where trainees must reach a predefined performance level before advancing to the next level of training. Senior clinicians typically conduct these assessments, which can be time consuming, subject to rater bias, and lack reproducibility. In the search for new technologies that are less dependent on clinicians, artificial intelligence (AI) has been promoted as a potential game changer for educational assessments. AI involves the use of computer programmes to perform tasks typically done by humans. AI can discover patterns in large and complex datasets without being explicitly programmed to do so. Using AI for assessment offers fast real-time feedback with high levels of reliability and has the potential to uncover previously unrecognised patterns associated with expert performance. However, AI assessments are narrow and transfer poorly across datasets, making them less versatile than clinical expert-based assessments. Additionally, the use of AI risks introducing systematic bias in assessment, which may reinforce social or racial inequality for certain groups of learners. One concerning aspect of AI assessments is that, unlike biased clinical expert-based assessments that may only affect a small number of individuals at a time, they have the potential to impact entire generations, ethnic groups or genders.

Despite the inherent limitations of AI assessments, there has been significant hype surrounding their use in medical education, as reflected in the numerous editorials and comments published on the topic in recent years. A recent review expressed concern that existing research on AI assessments in medical education remains under-theorised and under-conceptualised in terms of the use of contemporary validity frameworks. Claims are often made without empirical evidence to support the usability and validity of AI assessments. Furthermore, evaluations of AI assessments are often performed in isolation, which fail to uncover differences in the validity evidence supporting them compared with clinical expert-based assessments. For these reasons, many existing studies justify the use of AI for assessment of learning and performance but do not explore how, when and for what AI should be used.

To answer these knowledge gaps, we aimed to identify strengths and limitations of AI assessments (AIBA) compared with clinical expert-based assessments (EBA) under standardised and comparable conditions. We used Kane’s validity framework to systematically examine the weakest link in the validity argument underpinning the use of AIBA versus EBA of a complex clinical skill. In doing this, we sought to provide insights into how AI assessments should or should not be used in future practice as well as how to avoid pitfalls in the future development and use of AI in medical education. We identified the weakest link as the degree to which the assessments could differentiate between information-bearing and random patterns (detect a signal) on the level of individual scores and total scores.

2 | METHODS

The context of our study was assessment of a complex technical skill involving ultrasound-guided needle biopsy of the human placenta (chorionic villus sampling [CVS]). All performances were evaluated in the simulated setting. The study was conducted at the Copenhagen Academy for Medical Education and Simulation (CAMES) between September 2020 and October 2022. Ethical approval was obtained in terms of an exemption letter from the Ethical Committee of the Capital Region, Denmark (Protocol No. 19085543).

2.1 | Kane’s validity framework and analysis plan

We applied Kane’s validity framework to organise, evaluate and compare validity evidence for test scores provided by expert clinicians (expert-based assessment [EBA]) and an AI model (AI-based assessment [AIBA]).

The first step of Kane’s validity framework is to clearly state the intended interpretations and uses of the assessment. The construct of interest was competence in performing the CVS procedure. The intended use of the assessments was to evaluate performance of new trainees, to provide relevant feedback during training and to guide entrustment decisions in terms of progression from training in the simulated setting to the clinical setting.

The second step is to identify assumptions supporting the interpretations and uses and to organise them according to Kane’s four inferences: scoring, generalisation, extrapolation and implications. In total, we identified nine assumptions. The assumptions are summarised in Table 1. The list of assumptions acts as our hypothesis, also referred to as the interpretation and use argument (IUA).

The weakest and most questionable assumption should be prioritised first. Previous research has argued to follow a logical order from scoring to implication when no previous empirical evidence has been collected. Therefore, we prioritised to examine the plausibility of assumptions related to scoring (Assumptions 1–3), evidence of reproducibility (Assumption 4) and relation to different training levels (Assumption 6). The reason we prioritised these was that they relate to the degree to which a model can detect signal (discriminate between information-bearing and random patterns in the data). If this was not the case, then all remaining assumptions would be meaningless. A plan was made for the collection of validity evidence to support or refute the included assumptions by evaluating their respective plausibility. The results constitute our validity argument to support or refute the intended interpretations and uses of AIBA versus EBA, respectively. It should be noted that the validity argument of this study is not exhaustive for EBA or AIBA. The results aim to contrast the assessments methods and to indicate where the weak links remain in the validity argument for the two types of assessments.
<table>
<thead>
<tr>
<th>TABLE 1 Interpretation and use argument.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assumptions</strong></td>
</tr>
<tr>
<td>Scoring</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Generalisation</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Extrapolation</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Implications</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: AIBA, artificial intelligence-based assessments; EBA, expert-based assessments.

### 2.2 Participants

A sample of participants with different training levels was included: novices, intermediates and experts. The novices were medical students from the University of Copenhagen with a passed general anatomy exam and no formal experience in ultrasound. The intermediate group were ob-gyn trainees from university hospitals in Denmark with no prior experience in performing the CVS procedure but with obstetric ultrasound experience. The expert group included fetal medicine consultants from five university hospitals in Denmark with ample experience in performing the CVS procedure. The participants were recruited by e-mail and via fora on Facebook between September 2020 and January 2021. Written informed consent was obtained from all participants.

### 2.3 Equipment

A CVS manikin was used to simulate a pregnant woman of 12 weeks of gestation. The manikin consisted of four compartments: the abdominal wall; the uterus and amniotic cavity; the placenta; and a silicone model of a 12-week fetus. A sterile tray with relevant equipment was placed next to the manikin including one 20-ml syringe; one 15-cm 18 gauge biopsy needle; ultrasound gel; sterile swabs and one bowl with antiseptic; one specimen bowl; and sterile surgery cover. The participants were also supplied with sterile gloves. We used the GE HealthCare LOGIQTM e Ultrasound with a C1-5 RS probe and a GE HealthCare C1-5 non-sterile ultrasound needle guide. During all procedures, we recorded ultrasound output (Video 1) and two videos of the participants. One camera was placed in front of the participant to record hand movements and the sterile tray (Video 2) and the other on top of the ultrasound display to record head and eye movements (Video 3). Visualisations of the simulation set-up and video outputs are available in Appendix S1.

### 2.4 Cases

Participants were asked to complete two full CVS procedures including preparation of site and instruments, check for adequate sample and communication. All participants received a brief introduction to the manikin and the equipment. Novices and intermediates watched a video of a full procedure performed on a real patient and were provided a written step-by-step guide on how to perform the procedure prior to the test. Participants received a brief written case description. The two cases (i.e. Performance 1 and Performance 2) were identical apart from the placenta being placed in the left and right uterine wall, respectively.
2.5 | EBA

The clinicians were instructed and trained in using an assessment instrument that had been developed in a previous Delphi study involving international CVS experts.\textsuperscript{24} The final assessment instrument included 11 items that were scored using five-point Likert scales. In the context of our simulated setting, two of the items were omitted as they were irrelevant to the test set-up (sample preparation and handling, and documentation).

We recruited two fetal medicine consultants (E. T. and L. H.) to rate all the performances. They were recruited from a Swedish center for Fetal Medicine (Karolinska University Hospital, Stockholm, Sweden) to avoid that they would recognise the identity or charge of the participants (all Danish). The two raters were presented with outputs from Videos 1 and 2 (Appendix S1, Figure S3b). Participants’ faces were not visible to the raters, and the pitch of their voices was altered in a video-editing programme.\textsuperscript{25} Rater training was conducted prior to the assessments where the items were explained and discussed. Subsequently, the two raters independently rated two performances that were not included in the study of a novice and an intermediate. After each rating, they discussed their scores until obtaining consensus for each item.

2.6 | AIBA

The AIBA included five items selected by an educational technology engineer (M. B. S.), two fetal medicine consultants (O. B. P. and K. S.) and a medical education scientist (M. G. T.). The item selection was guided by the previous consensus study\textsuperscript{24} and technological feasibility.

Five items were selected based on\textsuperscript{24} free passage, eyes on screen, grip, needle in the image, and time. Item descriptions and definition of directories are provided in Table 4. Convolutional neural networks (CNN) were trained for each item separately. Data splitting was performed by randomly selecting 30–40 input images that were manually sorted into directories for each item and used for training of the CNN.

The training took place in Lobe (Microsoft). Thereafter, the models were exported and applied for each participant using Python (3.7.9) with libraries in OpenCV and NumPy.

After processing all recorded videos, principal component analysis (PCA) was used for dimensionality reduction of the five items into two dimensions (PC1 and PC2). K-means clustering was used to partition the test scores into two clusters. A linear decision boundary was defined between the two clusters defining two groups. The K-means cluster and group including most experts was labelled ‘expert’, and the other K-means cluster and group was labelled as ‘novice’. AIBA was calculated as the distance to the ‘expert’ K-means multiplied with factor +1 if the participant was grouped as ‘expert’ and −1 if the participant was grouped as ‘novice’ (Figure 1). Item scores to be used for feedback were defined as the number of true frames divided with total number of frames (Table 4). The model architecture is described in detail in Appendix S1.

2.7 | Statistics

EBA scores were converted to the percentage of maximum score (maximum EBA score = 30). Because the AIBA scores were not on an absolute scale, they were not converted.
AIBA items were weighted according to relevance using PCA. The explained variance was calculated for PC1 and PC2 with respective loadings. We calculated accuracy (number of correctly labelled images out of all images), precision (number of correctly labelled images out of all images labelled ‘true’), recall (number of correctly predicted images out of all images labelled ‘true’) and F1-scores (the harmonic mean of precision and recall) for each AIBA item to analyse and assess the four CNN models. 

One-way ANOVA was used to compare mean EBA item scores between the different training levels. EBA items that failed to discriminate were not considered supported by validity evidence and eliminated. We chose this approach for the EBA over the PCA to align its development and use with previous rater-based assessment instruments.26–28

Internal consistency (EBA) was calculated using Cronbach’s alpha. A G-study was conducted for EBA to estimate variance components for each facet (rater, item and case) and all potential interactions and to compute a G-coefficient. We estimated absolute and relative reliabilities for each facet. AIBA only includes one facet (case); thus, test–retest reliability was calculated using intraclass correlation coefficients, for single measures with an absolute agreement, two-way mixed-effects model.29 Both AIBA and EBA data were further analysed in a decision or d-study to understand how changing facet sampling impacted reliability. For the AIBA, this was consisted to increasing the number of cases analysed. The analyses were performed using EduG (Swiss Society for Research in Education Working Group).

To explore the relationship between the scores and different training levels, mean AIBA and EBA scores from Performances 1 and 2 were compared between the three training levels using a linear mixed-effect model to assess the main effect of repeated testing and interaction between training levels and testing. A repeated unstructured covariance structure was applied to account for the correlation of repeated effects and Šidák’s correction was applied for multiple comparisons. Insignificant effects were removed from the model. Pearson correlation coefficients were used to determine the correlation between mean EBA and mean AIBA scores.

Pass/fail levels were determined using the contrasting group method.26,30,31 with subsequent sensitivity and specificity analysis. A significance level of 0.05 was used throughout all analyses. All statistical analyses were conducted using IBM SPSS Statistics for Windows, Version 28.0 (IBM Corp, Armonk, NY, USA).

3 | RESULTS

A total of 45 individuals (11 experts, 12 intermediates and 22 novices) participated in the study. All participants completed two CVS procedures. One expert video was lost due to technical problems, and eight videos were obtained without sound, and therefore, not eligible for expert rating. Participant demographics are reported in Table 2. All test statistics are provided in Table 3, and mean scores for the two performances are provided in Figure 2. The nine assumptions in our interpretation and use argument are examined under each of the four inferences below.

3.1 | Evidence for scoring

3.1.1 | EBA

Items and scale anchors were selected and agreed on by a large expert panel using the Delphi method.24 Three out of nine items failed to discriminate between novices and experts (general preparations, F[2, 39] = 0.9, p = 0.44; identification and consent, F[2, 39] = 0.2, p = 0.81; and preparation of site and instruments, F[2, 39] = 1.3, p = 0.29) (Table 3). Internal consistency for the remaining six items was high (Cronbach’s α = 0.89). The six items reflected the procedure, ultrasound assessment before and after the procedure, communication skills and overall performance.

3.1.2 | AIBA

A multidisciplinary team selected feasible items from the Delphi study and converted them to Boolean (true/false) items eligible for AI assessment.24 Item scores were weighted according to relevance using PCA. PC1 and PC2 accounted together for 65.2% of the explained variance (PC1, PC2 [%]; 44.9, 20.3). Needle in the image (0.49) had the highest

| TABLE 2 Baseline demographics. |
|---|---|---|
| Participants | Experts | Intermediates | Novices |
| N | 11 | 12 | 22 |
| Age, years, mean [95% CI] | 53.6 [47.4, 59.9] | 34.1 [30.6, 37.5] | 25.0 [23.9, 26.0] |
| Female, n/N | 8/11 | 10/12 | 19/22 |
| Clinical experience, years, mean [95% CI] | | | |
| Fetal medicine consultant | 12.7 [6.3, 19.2] | — | — |
| Ob-gyn trainee | — | 2.2 [1.6, 2.8] | — |
| Number of performed CVS past year, n, mean [95% CI] | 81 [56, 106] | — | — |
| Number of performed AC past year, n, mean [95% CI] | 51 [34, 68] | — | — |

Abbreviations: CI, confidence interval; CVS, chorionic villus sampling.
TABLE 3 Test statistics for the three groups of participants across the two types of assessments (AIBA and EBA).

<table>
<thead>
<tr>
<th></th>
<th>Experts</th>
<th>Intermediates</th>
<th>Novices</th>
<th>p value</th>
<th>F(2, 38.4) =</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBA percentage of max score, Mean [95% CI]</td>
<td>83.8 [76.8, 90.8]</td>
<td>69.5 [62.1, 77.0]</td>
<td>54.5 [49.7, 59.3]</td>
<td>&lt;0.0001** &lt;b&gt; F(2, 38.4) = 25.5</td>
<td></td>
</tr>
<tr>
<td>Mean test difference [95% CI]</td>
<td>9.0 [3.1, 14.9]</td>
<td></td>
<td></td>
<td></td>
<td>F(1, 39.5) = 9.4</td>
</tr>
</tbody>
</table>

EBA item, mean (SD)

<table>
<thead>
<tr>
<th>Item</th>
<th>Experts</th>
<th>Intermediates</th>
<th>Novices</th>
<th>p value</th>
<th>F(2, 39) =</th>
</tr>
</thead>
<tbody>
<tr>
<td>General preparations</td>
<td>4.1 (0.6)</td>
<td>4.4 (0.7)</td>
<td>4.4 (0.8)</td>
<td>0.44*</td>
<td>F(2, 39) = 0.9</td>
</tr>
<tr>
<td>Identification and consent</td>
<td>4.9 (0.3)</td>
<td>4.8 (0.7)</td>
<td>4.9 (0.3)</td>
<td>0.81*</td>
<td>F(2, 39) = 0.2</td>
</tr>
<tr>
<td>Pre-procedure ultrasound assessment</td>
<td>4.2 (0.7)</td>
<td>3.5 (0.7)</td>
<td>3.0 (0.6)</td>
<td>&lt;0.01**</td>
<td>F(2, 39) = 14.3</td>
</tr>
<tr>
<td>Preparation of site and instruments</td>
<td>3.3 (0.9)</td>
<td>3.7 (1.2)</td>
<td>3.1 (0.8)</td>
<td>0.29*</td>
<td>F(2, 39) = 1.3</td>
</tr>
<tr>
<td>Selection of insertion site</td>
<td>4.2 (1.0)</td>
<td>3.4 (1.0)</td>
<td>2.7 (0.7)</td>
<td>&lt;0.01**</td>
<td>F(2, 39) = 10.7</td>
</tr>
<tr>
<td>Sampling technique</td>
<td>4.4 (0.6)</td>
<td>3.0 (1.2)</td>
<td>2.2 (0.7)</td>
<td>&lt;0.01**</td>
<td>F(2, 39) = 27.7</td>
</tr>
<tr>
<td>Post-procedure ultrasound assessment</td>
<td>4.1 (1.3)</td>
<td>3.9 (0.8)</td>
<td>3.1 (1.2)</td>
<td>0.04**</td>
<td>F(2, 39) = 3.6</td>
</tr>
<tr>
<td>Communication with patient</td>
<td>4.7 (0.5)</td>
<td>4.2 (0.4)</td>
<td>3.4 (0.7)</td>
<td>&lt;0.01**</td>
<td>F(2, 39) = 18.0</td>
</tr>
<tr>
<td>Overall assessment</td>
<td>3.9 (0.8)</td>
<td>2.9 (0.9)</td>
<td>2.1 (0.6)</td>
<td>&lt;0.01**</td>
<td>F(2, 39) = 22.4</td>
</tr>
</tbody>
</table>

AIBA scores, mean [95% CI]

<table>
<thead>
<tr>
<th></th>
<th>Experts</th>
<th>Intermediates</th>
<th>Novices</th>
<th>p value</th>
<th>F(2, 36.7) =</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean test difference [95% CI]</td>
<td>0.51 [0.35, 0.68]</td>
<td>-0.06 [-0.21, 0.10]</td>
<td>-0.20 [-0.32, 0.09]</td>
<td>&lt;0.01**</td>
<td>F(2, 36.7) = 28.7</td>
</tr>
</tbody>
</table>

Abbreviations: AIBA, artificial intelligence-based assessments; CI, confidence interval; EBA, expert-based assessments.

*One-way ANOVA comparing item scores.

**Linear mixed-effect model with repeated unstructured covariance structure.

*Significant p value.

3.2 | Evidence for generalisation

3.2.1 | EBA

The overall reliability was 0.59 (absolute) and 0.65 (relative). The generalisability analysis showed that case specificity was the most important determinant of reliability, i.e. increasing the number of cases had a greater effect on overall reliability than either increasing raters or items (Table 6). Inter-case reliability for a single rater was 0.44 and 0.49 (Table 5). Previous studies using similar assessments of ultrasound skills...
Table 4: Results of the AIBA item analysis including the performance of the CNN models and how they were weighted in the PCA (PCA loadings).

<table>
<thead>
<tr>
<th>AIBA item Description</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>PCA loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proportion of correctly labelled frames out of all frames</td>
<td>Number of correctly labelled frames out of all frames labelled ‘true’</td>
<td>Number of correctly predicted images out of all images labelled ‘true’</td>
<td>The harmonic mean of precision and recall</td>
<td>PC1</td>
</tr>
<tr>
<td>Free passage</td>
<td>0.72</td>
<td>0.88</td>
<td>0.54</td>
<td>0.78</td>
<td>0.44</td>
</tr>
<tr>
<td>Video frames (Video 2) were labelled ‘true’ if the needle was inserted parallel with the placenta and labelled as ‘false’ if the needle was inserted perpendicular to the placenta.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eyes on screen</td>
<td>0.87</td>
<td>0.93</td>
<td>0.82</td>
<td>0.87</td>
<td>0.25</td>
</tr>
<tr>
<td>Video frames (Video 3) were labelled ‘true’ if the participant looked on the screen and ‘false’ else.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grip</td>
<td>0.70</td>
<td>0.60</td>
<td>0.80</td>
<td>0.67</td>
<td>0.39</td>
</tr>
<tr>
<td>Video frames (Video 2) were labelled ‘true’ if the hand was placed around the syringe and the thumb was used to establish vacuum during aspiration and ‘false’ else.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Needle in image</td>
<td>0.79</td>
<td>0.69</td>
<td>0.92</td>
<td>0.79</td>
<td>0.49</td>
</tr>
<tr>
<td>Ultrasound images (Video 1) were labelled ‘true’ if the needle was visualised on ultrasound and ‘false’ else.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-0.60</td>
</tr>
<tr>
<td>Time from incision to extraction of the needle.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: AIBA, artificial intelligence-based assessments; CI, confidence interval; CNN, convolutional neural network; EBA, expert-based assessments; PCA, principal component analysis.

*Factor loadings of a variable quantifies the extent to which the variable is related to a given factor (PC1 or PC2). Factor loadings above 0.40 appear indicate a central factor and appear in bold.*
have shown good generalisation between contexts with reliability estimates that correspond to those obtained in our study.\textsuperscript{24,26,32–34}

3.2.2 | AIBA

The overall reliability with two cases was 0.61 (absolute) and 0.66 (relative). The reliability were 0.44/0.49 for a single case, 0.70/0.73 at three cases; 0.75/0.78 at four cases; and 0.79/0.83 at five cases. We tested the model on a different population in another dataset with a slight modification of the task to determine the generalisability and impact of data shift on the model. The population included 384 performances by 96 novices using the same simulator as used in the present study. The task was modified using multiple positions of the placenta and performed with or without a guide on the ultrasound probe/monitor. The robustness of the model was poor as it did not pick up signals for several of the data sources, in particular when analysing item grip. The model needed to be recalibrated to achieve the same level of performance as in the present study.

To assess if our simulation-based test set-up was sufficient to establish a reliable measure of participants’ skills in performing the CVS procedure, we compared the content of our test set-up with curricular content deemed relevant by an international panel of CVS experts.\textsuperscript{24} Our test set-up represented three out of three major topics listed under “Technical skills: ultrasound” and three out of six major topics listed under “Technical skills: sampling”. Technical skills, such
as free hands technique, single- versus double-handed procedures and technically challenging patients, were not represented in our test set-up.

### 3.3 Evidence for extrapolation

#### 3.3.1 EBA

There was a significant effect of level of experience on scores ($F_{[2, 38.4]} = 25.5, p < 0.01$). The pairwise comparison demonstrated a significant difference between experts and novices (mean percent difference [95% CI] = 29.8 [19.4–40.1], $p < 0.01$), experts and intermediates (mean percent difference [95% CI] = 14.6 [2.0, 27.1], $p = 0.02$), and intermediates and novices (mean percent difference [95% CI] = 15.3 [4.4, 26.1], $p < 0.01$).

#### 3.3.2 AIBA

There was a significant effect of the level of experience on AIBA scores ($F_{[2, 36.7]} = 28.7, p < 0.0001$). The pairwise comparison demonstrated a significant difference between experts and novices (mean score difference [95% CI] = 0.72 [0.48, 0.95], $p < 0.01$) and experts and intermediates (mean score difference [95% CI] = 0.57 [0.30, 0.84], $p < 0.01$). However, there was no significant difference between intermediates and novices (mean score difference [95% CI] = 0.14 [–0.08, 0.37], $p = 0.31$).

### 3.4 Evidence for implications

#### 3.4.1 EBA

A pass/fail level was determined at a total score of 21 points corresponding to 69.3% of the max score (Figure 5a). The sensitivity and specificity for classifying experts and novices using this threshold were 0.80 and 0.96, respectively. One novice passed the test (false positive), and two experts failed the test (false negative).

#### 3.4.2 AIBA

A pass/fail score was determined at 0.07 (Figure 5b). The corresponding sensitivity and specificity were 0.80 and 0.96, respectively. One novice passed the test (false positive), and two experts failed the test (false negative).

Previous studies have shown the usefulness of AI-generated feedback; however, the choice of AI measures affect its usability. This study did not collect any data on the impact of either EBA or AIBA feedback on learning outcomes. The input used for the
assessment method affects what type of feedback they enable. The EBA generated assessments over a wider range of domains compared with AIBA (Figure 3); three out of four AIBA items (eyes on screen, grip and needle in the image) could be mapped under a single EBA item (sampling technique). Also, although the EBA provided readily interpretable explanations for the final scores that could be used for feedback, the AIBA did not provide interpretable explanations for its predictions due to the highly non-linear and complex analyses used for the final model predictions.

Appendix S2 provides a correlation matrix between the AIBA and EBA items demonstrating that while mean scores are correlated, item scores are not. This may indicate that the same underlying construct is measured drawing on different sources of information, potentially offering complementary benefits.

4 | DISCUSSION

We prioritised and tested nine assumptions relating to the interpretation and use of assessments made by an AI model (AIBA) and trained clinical experts (EBA). We organised and collected validity evidence according to Kane’s four inferences categories and found several important differences between AIBA and EBA.

First, validity evidence to support the assumption that the assessments reflect the observed performance demonstrated a construct underrepresentation in AIBA compared with EBA. The holistic approach of clinical EBA is difficult to incorporate in AI assessments. Rather than observing behaviours, AI assessments are inherently limited to what can be measured. The complexity of an assessment may thereby be reduced into narrow technical measurements with little consideration for what is relevant for the construct of interest (e.g. CVS-competence). With that said, for specific uses, a narrow but accurate AI assessment might still be preferred when it comes to specific technical skills such as visualisation of the needle on ultrasound or hand–eye coordination because of the high reproducibility and consistency. However, if assessment is focusing too narrowly on specific aspects and ignoring overall construct representation, it risks losing its value and, in the end, becomes merely a stopwatch.

Second, AIBA did not provide actionable feedback or explanations of assessment scores to the same extent as EBA—a finding that corroborates existing concerns expressed in the AI literature around the need for explainable AI for clinical decision support. One failure with the AIBA scores was lack of insights into the neural network and PCA decisions on a technical level. Previous studies have made attempts to increase the level of explainability by introducing heat maps (saliency maps highlighting pixels that are of importance to the AI decision) or other post hoc explanation techniques. However, they have been criticised for limited usability and for providing unstable output. Consequently, existing AI approaches often fail to offer direct insights into behaviours associated with the development of competence.

Third, the overall reliability of EBA and AIBA was comparable, demonstrating that case specificity was the most important determinant of reliability. Whereas both EBA and AIBA could discriminate between different training levels, the K-means grouping penalisation built into the AIBA model entailed stronger discrimination between intermediates and experts. In contrast, EBA discriminated better between intermediates and novices, suggesting that the assessment methods performed differentially depending on the learner level. As for binary expertise classification, AIBA and EBA were equally efficient. Although some expert performances only passed AIBA or EBA, no expert performances failed both the AIBA and EBA. When evaluating item-based comparisons across AIBA and EBA, different types of information were afforded, which may support the notion that the
assessments offered complimentary information on the same underlying construct. Recent literature has called for AI assessments that support rather than mimic the work of clinicians.\textsuperscript{5} As such, combining AIBA and EBA scores into a composite score has the potential to reduce test bias and increase reliability above the level of the individually best measures.\textsuperscript{11} This approach may be used in the future to improve reliability of assessments similar to a ‘double read’ as known from the imaging specialties. This would overcome some of the proposed limitations of AIBA (lack of explainability and lack of robustness) and EBA (inconsistency, flaws due to lack of attention and rater fatigue) and instead create more robust, explainable and reliable assessment systems.

Investigating validity is an ongoing process. The aim of our paper was not to provide a full validation of any single assessment but rather to contrast the two assessment methods and illustrate how Kane’s validity framework can be useful in doing so. Our study provides supportive evidence for assumptions referring to the scoring and generalisation inference for EBA. As for AIBA, the weakest link in the validity argument was the poor robustness where further training of the model and evidence to support its robustness are needed. In addition, more evidence must be collected for both AIBA and EBA to support the use of scores to guide entrustment decisions where the relation to measures such as clinical performance should be investigated. Following previous validity reports on AI assessments, accuracy, precision, sensitivity and F1-score were reported in this study to support the scoring inference.\textsuperscript{42,43} However, in accordance with contemporary guidelines for trustworthy AI in medicine,\textsuperscript{44} our study emphasises the importance of reporting validity threats in terms of robustness and explainability.

The strengths of this study include the standardised context of data collection and the use of a contemporary validity framework to organise, prioritise and compare validity evidence. Kane’s validity framework focuses on the link between an assumption and the evidence to support it rather than specific measures of validity. That makes Kane’s framework eligible for AI assessment where traditional validity sources are not always applicable.\textsuperscript{22}

Our study also has some limitations. The content of EBA and AIBA were both based on the recommendations from a previous consensus study. However, although the content of EBA could be directly applied, the AIBA format did not allow for the same type of observations. To select items for AIBA, a multidisciplinary team evaluated how consensus recommendations could be turned into observable features. By doing this, we introduced a systematic difference. Yet the purpose of our study was not to avoid these differences but rather to highlight how the structured use of a validity framework for comparisons of AI and clinician-based assessment offers different insights to their weakest links. By comparing where there is similar validity evidence and where the EBA or AI approach offers incremental validity, programmes of assessment can intelligently combine clinician and algorithmic input to make high-quality decisions.

Although the sample size of this study corresponds to previous validation studies within the domain of health professional education research,\textsuperscript{1,3,45} it would be considered small from an AI perspective. This has implications for the interpretation of our study results as we may underestimate the potential value of AI-based assessments. The robustness of the AIBA model may be improved by including more diverse training sets and by adjusting the model’s hyperparameters to avoid overfitting to the training dataset.\textsuperscript{46–49} However, the need for cross-validation and very large datasets may ultimately hinder the accessibility and use of AI for assessment purposes, in particular, when compared with EBA that work after minimal rater instruction.

5 | CONCLUSION

Construct underrepresentation, lack of explainability and threats to robustness were identified as weak links in the use of AIBA compared with EBA. Our findings suggest that combining AI and clinical expert-based assessments may offer complementary benefits. However, it is important to note that significant efforts are required to calibrate AI models when using them for slightly different datasets, populations or tasks.

AUTHOR CONTRIBUTIONS

Vilma Johnsson et al. substantially contributed to the conception and design of the work. She has conducted the analysis and interpretation of data. She has drafted the work and provided final approval of the version to be published. She has revised the work critically for important intellectual content and provided final approval of the version to be published. He agrees to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Morten Bo Søndergaard has substantially contributed to the conception and design of the work. He has conducted analyses and interpretation of data. He has revised the work critically for important intellectual content and provided final approval of the version to be published. He agrees to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Karin Sundberg has substantially contributed to the conception and design of the work. He has revised the work critically for important intellectual content and provided final approval of the version to be published. He agrees to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Eleonor Tiblad has conducted analyses of data. She has revised the work critically for important intellectual content and...
provided final approval of the version to be published. She agrees to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Olav Bjørn Petersen has substantially contributed to the conception and design of the work. He has revised the work critically for important intellectual content and provided final approval of the version to be published. He agrees to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Martin G. Tolsgaard has substantially contributed to the conception and design of the work. He has conducted analyses and interpretation of data. He has substantially contributed to the drafting of the manuscript and revised the work critically for important intellectual content. He provided final approval of the version to be published. He agrees to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGEMENTS
The authors extend their sincere gratitude to the medical students, ob-gyn trainees and fetal medicine consultants who enthusiastically participated in this study.

CONFLICT OF INTEREST STATEMENT
The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT
Ethical approval was obtained in terms of an exemption letter from ETHICS STATEMENT available due to privacy or ethical restrictions. The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

REFERENCES


SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.