Developing a consensus-based scoring rubric to enhance practice-based assessment of student nurses' clinical competence: a Delphi study

Authors:

Dr Ibraheim Almalkawi (London South Bank University) (Corresponding author)

Professor Rebecca Jester (University of Wolverhampton)

Dr Louise Terry (London South Bank University)

ABSTRACT

Background: Concerns about reliability and validity of practice-based assessment of professional competencies are frequently reported in the literature. Difficulty in understanding competency statements or distinguishing different achievement levels has been found to be a major factor.

Objectives: To develop a consensus-based scoring rubric based on stakeholders’ interpretations of level descriptors for student nurses’ professional values competencies.

Design: Two rounds of Classic e-Delphi

Settings: This study was conducted in a London based university using Bristol Online Survey website as a host.

Participants: 100 stakeholders with vested interests in undergraduate pre-registration nurse education were purposefully invited to participate.

Method: Round one collected free-text interpretations of the United Kingdom Nursing and Midwifery Council professional values competency statements. Round two used a Likert scale questionnaire to measure the level of agreement to the level descriptor statements generated through round one. Responses were analysed through content analysis in round one and consensus measure in round two. A threshold of 70% agreement to determine consensus was set in advance.

Results: In round one, 47 participants provided their interpretations of the competency statements. In round two, 51 participants completed the questionnaire. All 24 items achieved a strong consensus with 86%-100% of participants agreeing or strongly agreeing with the statements.

Conclusions: A Delphi study was successfully used to develop a consensus-based scoring rubric with clearly stated descriptors for professional values competency statements. This scoring rubric holds the potential to enhance practice-based assessment across all healthcare professional disciplines.

## KEY WORDS

Student nurses

Practice-based assessment

Level of competence

Interpretation

Professional values

Delphi

Rubric

## INTRODUCTION

The complexity of assessing clinical practice for all healthcare professionals is a long-standing issue and has received much attention for decades. Internationally, there are concerns that practice-based assessors lack confidence in carrying out their role, feel unprepared and unwilling to fail students (Hunt et al., 2012; Trede and Smith 2012; De Vos et al., 2019; Hishiti et al., 2020), which places patients at risk (Duffy, 2003; Cassidy et al., 2017). Ambiguous language used in Practice Assessment Documents (PADs) and difficulty in identifying the benchmarks that discriminate between levels of practice have been identified as significant barriers to effective practice-based assessment across health professions (Butler et al., 2011; Miller, 2010; Almalkawi et al., 2018).

In the United Kingdom (UK), the Higher Education Institutions (HEIs) in the London region have used the Nursing and Midwifery Council (NMC)-approved single Pan London Practice Assessment Document (PLPAD) to provide uniformity of the PADs used in clinical practice. The NMC Standards for Student Supervision and Assessment (NMC, 2018) have introduced a tripartite model comprising practice supervisors, practice assessors and academic assessors to evaluate and confirm students’ proficiencies. However, the practice assessor role will continue to be performed by registered nurses based on the practice supervisors’ feedback (NMC, 2018). Considering that Practice assessors may spend little clinical time with students and will be reliant on the feedback from practice supervisors to inform their decisions, there is a clear need to find ways to develop transparent and common language to interpret different levels of competence to improve the validity and reliability of practice assessments (Cassidy et al., 2020).

## BACKGROUND

Many studies raise concerns related to the lack of reliability and validity of assessment in practice-based placements including medicine (Govaerts et al., 2013), occupational therapy (Nicola-Richmond, 2017), social work (Finch, 2017), pharmacy (Croft et al., 2019) and radiography (Hishiti et al., 2020). Nursing shares the same concerns that mentors’ judgements are subjective and do not always accurately reflect students’ performance (Cassidy et al., 2017; Nugent et al., 2020). There is reluctance to fail underperforming students in practice settings (Duffy, 2003; Larocque and Luhanga, 2013; Hughes et al., 2020). Internationally, there is evidence of similar concerns reported in Australia (Miller, 2010), Canada (Yonge et al., 2011), Finland (Jokelainen et al., 2013), Ireland (Kennedy and Chesser-Smyth, 2017), New Zealand (Gallagher et al., 2012), Norway (Bachmann et al., 2019) and the United States of America (DeBrew and Lewallen, 2014).

The key factor is that assessors lack understanding of competency statements regarding the appropriate level of competence (Almalkawi et al., 2018). One of the factors contributing to the ‘failing to fail’ phenomenon relates to the ambiguous terminology of the PADs as the language used in the documents is described as vague and containing too much academic jargon (Neary, 2001; Cassidy et al., 2017; Almalkawi et al., 2018; Bradshaw et al., 2019).

Consequently, mentors experience problems translating and applying assessment outcomes into observable practice activities. Mentors and students have reported spending significant time trying to work out what the competency statements mean rather than assessing the student against them (Neary, 2001). In particular, studies show that mentors find the ‘softer’ aspects of competence such as attitudes and behaviours to be notoriously challenging to define, measure or assess (Hunt, 2014; Strauss, 2016). Thus, when required to justify decisions regarding students not meeting competency standards, mentors struggle to prove their concerns are justified, feel unprepared and become unwilling to fail students (Duffy, 2003; Brown et al., 2012; Bradshaw, 2019).

A link between mentors’ reluctance to fail underperforming students and difficulties in identifying the benchmark of what constitutes a pass or a fail is acknowledged in the literature (Heaslip and Scammell, 2012; Cassidy et al., 2017; Almalkawi et al., 2018). Calman et al., (2002) identified that there was no agreed strategy in nursing for describing performance at different levels. The lack of transparent criteria against which students’ competence can be judged not only influences the accuracy of understanding and measuring levels of competence, but also how mentors deliver effective and constructive feedback (Fitzgerald et al., 2010; Almalkawi et al., 2018). While there are a variety of assessment taxonomies in existence, ambiguous language and difficulty in identifying performance levels continue to be problematic (Almalkawi et al., 2018). Significant amounts of literature have debated the appropriateness and quality of existing assessment tools (Watson et al., 2002, Immonen et al., 2019; Cassidy et al., 2020), yet have produced no resolution to the problem of separating the competent from the not-yet-competent and the incompetent.

An integrative review carried out prior to the study presented here highlighted the need to establish a transparent and common language to interpret different levels of competence to help assessors define what is expected of students and for students to identify what they are expected to achieve (Authors, 2018). The review established that central to the problem of practice-based assessment not being fit for purpose is a continuing inconsistency and difficulty in interpreting levels of performance when assessing nursing students in practice as well as difficulty in providing constructive feedback for formative learning (Authors, 2018).

Several researchers have suggested conducting research to explore whether a scoring rubric, with a transparent and common language to interpret different levels of competence, offers a solution to the challenges faced in practice-based assessment (Donaldson and Gray, 2012; Heaslip and Scammell, 2012; Shipman et al., 2012; Frentsos, 2013). This and the authors’ own experience inspired this current study.

THEORETICAL FRAMEWORK

The study design was influenced by two complementary theoretical frameworks: 1) assessment *for* learning and 2) authentic assessment. Assessment *for* learning recognises the power of formative assessment and feedback on driving learning (Schuwirth and Van der Vleuten, 2011). This is achieved through understanding the criteria that embody what it means to do well, shifting the focus towards learning, giving students opportunities to test-out subject knowledge and rehearse relevant skills before these are summatively assessed (McDowell et al., 2011). Oyinloye and Imenda (2019) found that assessment *for* learning strategies empowered students to take active part in their own learning and in developing self-assessment abilities.

Authentic assessment, advocated by Wiggins (1989), refers to situational or contextual realism of the proposed tasks and involves assessing the application of process and product in the real world. By taking into account the actual performances the profession wants students to achieve, criteria can be designed to replicate the actual challenges facing students in the real world, helping them to problem-solve, think critically and handle complex tasks (Wiggins, 1989). Authentic assessment seeks students’ application of knowledge to evaluate and address real problems (Wiewiora and Kowalkiewicz, 2019), prepare them to take responsibility for *who* they are becoming, rather than simply replicating practices without thinking (Ajjawi et al., 2020).

What underpins assessment *for* learning and authentic assessment is that the criteria are known in advance to focus both the student and the assessor. Central to assessment *for* learning and authentic assessment characteristics is the use of explicit and transparent criteria that describe what the target performance that should be achieved looks like, enabling students and assessors to make formative assessments and continue to modify practice-based educational opportunities to reach their full potential (Montgomery, 2002; Haines et al., 2013). Winkelbauer (2020) supported this argument stating that authentic assessment is effective in changing practice and, subsequently, improves patient care. In the literature, authentic assessment is frequently associated with the use of scoring rubrics. Boateng et al. (2009) reported that scoring rubrics provide a learner-centred assessment approach that focuses on encouraging behavioural change in learners. They promote consistency, encourage self-assessment, provide timely feedback, hence, provide a potential solution to the subjectivity of performance-based assessments in clinical settings.

## AIM OF THE STUDY

The aim of the study was to develop a consensus-based scoring rubric based on stakeholders’ interpretations of level descriptors for undergraduate student nurses’ professional values competencies in the PAD.

## STUDY DESIGN

The study adopted a two-round ‘classic e-Delphi’ presented in Figure 1. In keeping with the theoretical underpinning of assessment *for* learning and authentic assessment, it was crucial to select a method that ensured interpretations of the competency statements and development of level descriptors that were unambiguous, free from academic jargon, reflected the complexity of real life and supported by relevant stakeholders.

*(Insert Figure 1 here)*

Questionnaires were used in each of the two rounds, which used Bristol Online Survey (BOS) website as a host. The first round gathered stakeholders’ interpretations of the competency statements, where a qualitative approach using free-text questions was adopted. The second round used data from the first round to formulate attitudinal statements to establish the degree of consensus or divergence between the stakeholders’ interpretations of the level descriptors.

SAMPLE/PARTICIPANTS

Creating the expert panel meant recruiting a heterogeneous sample to ensure a broad range of stakeholders’ perspectives were sought. Representativeness, according to Keeney et al., (2011), is based on the qualities of the expert panel therefore sampling techniques such as purposeful, convenience and snowballing are appropriate. Bowling, (2014) advised that purposive sampling is appropriate where respondents are selected because they have knowledge that is valuable to the research. Therefore, in order to include a diverse but representative sample in this study, a purposive sampling approach was selected to recruit a group with particular characteristics. The definition of ‘expert’ in Delphi methods remains ambiguous and includes: informed individual (Mckenna, 1994), specialist in the field (Goodman, 1987), or knowledgeable about the subject (Lemmer, 1998). The most fitting definition of the term ‘expert’ was suggested by Keeney et al., (2011, p. 24) as “anyone with relevant input”. Guided by this definition, the criteria for inclusion in this study entailed all participants who have knowledge and experience in assessing the practice-based component of pre-registration education, being assessed or having a vested interest in the process, as well as having sufficient time and willingness to participate. Five groups of stakeholders were identified: practice-based assessors, nursing students, nurse educators, service users and strategy/policy shapers. Pre-determined selection criteria specific to each group is summarised in Table 1. Participants remained anonymous; they were only identified by their stakeholder group.

*(Insert Table 1 here)*

High attrition rates are frequently associated with Delphi in the literature (Donohoe and Needham, 2009) which can lead to response bias (Keeney et al., 2011). Several strategies were employed to enhance response rates and reduce attrition (Table 2).

*(Insert Table 2 here)*

When deciding the sample size for this study, considerations were given to the importance of recruiting a heterogonous sample to ensure a spectrum of stakeholder opinions was presented and the size was large enough to mitigate for the risk of attrition (Donohoe and Needham, 2009). Equally, the data generated needed to be manageable as large panels make analysis more time-consuming with limited benefits (Skulmoski *et al.,* 2007). Therefore, 100 potential participants were invited (twenty per group) with the aim of recruiting ten participants from each of the five groups shown in Table 1.

DATA COLLECTION

The two rounds were conducted between September - December 2015. Round one adopted the classic e-Delphi approach comprising a qualitative open-ended round, which ensured participants were given complete freedom to express themselves in their own words and thereby identify comprehensive possibilities (Hasson et al., 2000). Participants were asked to write their own descriptors to reflect their understanding of the eight professional attitude, behaviour and responsibility statements in the PLPAD (Table 3). To benchmark each performance level, the ‘Successful Outcomes Markers’ model comprising the headings ‘Don't want to see’, ‘Expect to see’ and ‘Love to see’ was used (Burkhart-Kriesel et al., 2011).

*(Insert Table 3 here)*

The questionnaire for round two aimed to gauge stakeholders’ agreement to the level descriptors developed in round one. Due to the anonymity, it was not possible to identify who responded in round one, therefore, invitation emails were sent to all those who had been previously invited. The data in round two were ordinal to enable the variables to be ranked on a scale of increasing magnitude.

A neutral option was included in the questionnaire to allow those who genuinely had no opinion to indicate this **(**Gray et al., 2017). Every participant was presented with all eight statements, each in turn being presented in a screen with the statements provided at the top of the webpage and below it the three level attributes were provided. Across each level, an attitudinal five-point Likert scale was positioned ranging from strongly agree to strongly disagree, commonly used to collect this type of data (Wood et al., 2011). In addition, participants were given the opportunity to rate their preferred benchmarking style among the commonly used level descriptor methods, these are:

* Don't want to see - Expect to see - Love to see.
* Developing - Competent – Exemplary.
* Does not meet expectations - Meets expectations - Exceeds expectations.
* Not achieved - Achieved – Merit.
* 1 - 2 - 3 (Numerical where 2 is needed for a pass).

The questionnaires for both rounds were piloted on representative samples resulting in shortening the questionnaire in round one and minor wording changes for questionnaires in both rounds.

ETHICAL CONSIDERATIONS

Ethical approval was granted by the University Research Ethics Committee. The Chair of the Pan London Practice Learning Group granted permission to use the professional values statements. Consent was implied by completion of the BOS. Reponses were anonymous so only stakeholder group was identifiable.

DATA ANALYSIS

Qualitative content analysis (QCA) was employed for the initial unstructured questionnaire in round 1 to identify and group statements generated by the panel (Keeney et al., 2011). QCA allows the researcher to pull out key words and statements that have similar meanings then group and collapse them into one statement, while ensuring that the statements provided by the panel remain as true to the wording as possible (Keeney et al., 2011). Deductive content analysis was used to amalgamate all descriptors for each category to form single statements. These were consolidated into a scoring rubric (an example of which is provided in Table 4).

*(Insert Table 4 here)*

The data analysis in round two aimed to measure the level of consensus. In the absence of an agreed threshold to determine a cut-off point for reaching consensus (Keeney et al., 2011), consensus was considered achieved in this study when agree or strongly agree responses scored 70% or higher, reported by Diamond et al. (2014) as a generally accepted benchmark. Free-text comments in round two were analysed and taken into consideration in the scorning rubric’s final version.

VALIDITY AND RELIABILITY

Problems with reliability of the Delphi method are commonly reported in the literature and predominantly relate to the variations of the procedural applications of the method, which makes standardisation of the method to evaluate its reliability difficult (Keeney et al., 2011). In light of potential reliability shortcomings, Hill and Fowles (1975) suggested a set of considerations to enhance methodological rigour in Delphi. Table 5 summarises these considerations including how they were addressed in this study.

*(Insert Table 5 here)*

Hill and Fowles (1975) claimed that validity is enhanced in Delphi given that the method is based on a group opinion, and the decisions are then strengthened by the reiteration and controlled feedback process to challenge assumptions. Delphi provides content and face validity based on confirmative judgements of a panel that have knowledge and interest in the topic (Keeney et al., 2011).

Hasson et al. (2000) argued that concurrent validity is achieved through reaching a consensus using successive rounds of the questionnaires, and predictive validity is measured in terms of accuracy of the Delphi (Von der Gracht, 2008), and is proof of the method validity (Keeney et al., 2011). Validity is ultimately affected by response rates (Hasson et al., 2000). In this study, representation across all stakeholder groups was achieved with 47% and 51% of invitees participating in rounds 1 and 2 respectively. There was a slight difference in the number of participants between the two rounds, but since they had shared stakeholder characteristics and came from the same groups that were originally invited to participate, this difference was unlikely to have an impact on the consensus achieved.

Holloway and Wheeler (1996) believe that Delphi overlaps both quantitative and qualitative approaches, so advocate using the term ‘trustworthiness’ to determine the appropriateness and effectiveness of a Delphi. Trustworthiness, according to Lincoln and Guba (1985), comprises credibility, dependability, confirmability and transferability. As this study adopted a mixed methods approach, efforts were made to achieve trustworthiness. Recruiting a sample representative of stakeholders involved in nurse education enhanced credibility. Dependability was also enhanced by the controlled feedback process. Confirmability was achieved by the researchers maintaining an ‘outsider observer’ to ensure that the findings were shaped by the participants views. Transferability was also enhanced by describing in detail the study procedure, findings and interpretation of the data.

RESULTS

**Round 1**

Following 100 invitations, 47 questionnaires were returned in round 1(a response rate of 47%). There was a good spread of responses representative of all groups as shown in Figure 2.

*(Insert Figure 2 here)*

A total of 24 attributes (three level descriptors per statement) were drafted into a scoring rubric ready for round 2 to seek verification from the panel on their level of agreement to the interpretations.

**Round 2**

The same Delphi panel members from the first round were asked to rate their level of agreement to the 24 attributes in the scoring rubric. Of the 100 invitations 51 participants completed the questionnaire (a response rate of 51%). As with Round 1, there was good representativeness from across all stakeholders’ groups (Figure 3).

*(Insert Figure 3 here)*

All 24 items in the questionnaire achieved a strong consensus with 86%-100% of participants agreeing or strongly agreeing with the statements.

The agreed consensus threshold of 70% in the interpretations of all 24 statements was exceeded in round 2, the need for further rounds was eliminated. The 24 statements can be viewed within the finalised scoring rubric (Table 6).

*(Insert Table 6 here)*

Participants were given the opportunity to expand on the fixed choice responses using the free-text box provided for making additional comments. These comments provided a unique opportunity for participants to evaluate further the contents of the scoring rubric. In total, 46 comments were recorded. Comments on how valuable participants felt the rubric would be in relation to addressing the complexity and challenge of practice-based assessment included ‘very clear guidance’ and ‘clearest set of values so far’. Some commented on terminology such as ‘replace the word scruffy with unkempt’. Others suggested adding words like ‘actively’ to ‘students seek to understand…’ or advised sub-dividing a descriptor sentence into two shorter sentences. Some merely clarified responses. Where appropriate, these comments were taken into consideration in the finalised scoring rubric as seen in Table 6.

Participants were also asked to rate their preferred terminology to use for benchmarking. Most participants (84%) strongly favoured or favoured to use ‘Does not meet expectations’ – ‘Meets expectations’ – ‘Exceeds expectations’, which was adopted in the finalised draft of the scoring rubric.

DISCUSSION

Within the context of this study, the Delphi consensus method was selected to encapsulate the insights of different stakeholders to arrive at an informed group consensus in gaining a collective interpretation of three level descriptors for the professional attitude, behaviour and responsibly statements. This study was successful in reaching a very strong consensus (ranging from 86% to 100%) on shared interpretations of level descriptors for the professional values competency statements within the PLPAD. Delphi, as a consensus method, offers a highly structured and effective approach in facilitating democratic deliberation among geographically dispersed stakeholders with various roles who have vested interests in the outcomes (competent nurses) to apply their collective judgement. Additionally, a collaborative approach that involves mentors and students in developing tools is essential for success in improving practice-based assessment (Heaslip and Scammell, 2012; Cockett and Jackson 2018; Cassidy et al., 2020).

These shared understandings informed the construction of an original scoring rubric with transparent and unambiguous language to interpret different levels of competence. This consensus-based scoring rubric has the potential to overcome the difficulties in understanding the language used or in identifying the appropriate level of competence, reported in the literature to be a significant barrier to the reliability and validity of assessment in practice placements (Neary, 2001; Cassidy et al., 2017; Almalkawi et al., 2018). The transparency provided through making the scoring rubric available to both mentors and students in advance resonates with the concepts of assessment *for* learning, emphasising both process and product in practice-based assessment, which stresses the formative potential of the assessment through individualised feedback on performance (Oyinloye and Imenda, 2019). It could help mentors define what is expected of students and for students to identify what they are expected to achieve (Gray and Donaldson, 2009; Shipman et al., 2012; Frentsos, 2013). The characteristics of the scoring rubric also resonate with the concept of authentic assessment by employing ‘real world’ tasks that mutually engage students and mentors in collaborative and meaningful activities (Winkelbauer, 2020).

Furthermore, the Delphi as a method was suitable to determine if consensus exists in interpreting professional attitude and behaviour competency statements. The Delphi method has gradually found its place in academic studies, becoming highly popular in the early 21st century (Landeta, 2006). It is described as “a method used to systematically combine expert knowledge and opinion to arrive at an informed group consensus on a complex problem” (Donohoe and Needham, 2009, p. 416).

In response to the literature reporting concerns about practice-based assessors’ reliability and validity in assessing clinical competence and difficulties in interpreting different levels of competency statements in practice documents which is a major factor affecting the reliability and validity of assessments (Almalkawi et al., 2018; Bradshaw et al., 2019), this study provides evidence that clear and unambiguous language to describe competency statements that distinguish between different levels of competence is achievable with the use of a well-designed consensus-based scoring rubric. It also shows that the problem of students being educated to meet minimum standards, reported by McCarthy and Murphy (2008) as occurring in the absence of recognisable standards that indicate students’ levels of performance, can be overcome.

In this study, out of the 100 invitations, the response rate in rounds I and 2 was 47% and 51% respectively. There are no criteria for acceptable response rates or attrition in Delphi. The literature reports response rates ranging from 8% (Cooney et al., 1995) to 100% (Owens et al., 2008), with some authors recommending a 70% response rate to maintain rigour (Bork, 1993; Sumison, 1998). However, achieving 70% requires considerable effort where the researcher must know the identity of respondents, and non-respondents must be pursued, which presents problems with anonymity (Keeney et al., 2011). Nonetheless, the consensus reached in this study was based on a good spread of responses representative of all groups shown in Table 1. Having the opportunity to identify and agree the components, the consensus reached among the group can be considered as achieving face, content and concurrent validity (Keeney et al., 2011).

The panel was provided with the commonly used level descriptors. Contrary to the theoretical presumption that the Successful Outcome Markers approach of ‘Don't want to see’, ‘Expect to see’ and ‘Love to see’ would be ranked high, most participants preferred to use ‘Does not meet expectations’ – ‘Meets expectations’ – ‘Exceeds expectations’, which was adopted in the final rubric draft. There were no comments provided by the panel members to explain their choices but having the word ‘expectations’ may have provided the panel with clearer distinction of the levels. Here, Delphi is utilised as a social democratic procedure that emphasises social involvement in the development of a scoring rubric. This addressed the problem reported by Neary (2001, p. 4) that mentors claim that assessment strategies are “too objectively bound” and designed to suit the HEI academic requirements rather than providing sufficient scope to reflect students’ performance accurately.

LIMITATIONS

Although this study drew on a national sample of stakeholders, the generalisability of the findings may be affected by recruiting student nurses from one university and most of the panel members were based in London (using the PLPAD), which was necessary as the PLPAD was used in the London region at the time of conducting the study. Therefore, the findings have some limitations on generalisability beyond the single London HEI and further research with a wider nationally-representative sample is needed. While the purposive sampling method used in both phases was necessary to reach the target population, it is also considered a limitation of this study due to potential to introduce bias in the sampling process (Sharma, 2017).

There were some respondents who disagreed or strongly disagreed with the few statements in the scoring rubric, but there were no comments provided by participants in the free-text box to explain the reasons. Therefore, not knowing whether respondents did not agree with the statements or they misinterpreted the direction of the Likert scale is a limitation.

CONCLUSION

This study is the first to use a Delphi consensus method to elicit stakeholders’ interpretation of the professional attitude, behaviour and responsibility statements in the PLPAD and subsequently design and develop a consensus-based scoring rubric to improve the validity and reliability of practice assessment decisions. This study was successful in reaching a very strong consensus (ranging from 86% to 100%) on shared interpretations of level descriptors for professional attitude and behaviour competency statements, which is a unique outcome of this study. The scoring rubric presented here for professional attitude, behaviour and responsibility competency statements holds potential for use across all practice-based healthcare professional disciplines. A feasibility study to evaluate the usefulness of the consensus-based scoring rubric in enhancing the quality of the formative and summative assessment of nursing students from assessors’ and students’ perspectives has since been conducted and will be reported separately.

REFERENCES

Ajjawi, R., Tai, J., Huu Nghia, T.L., Boud, D., Johnson, L., Patrick, C.J., 2020. Aligning assessment with the needs of work-integrated learning: The challenge of authentic assessment in a complex context. Assessment and Evaluation in Higher Education, 45(2), pp.304-316. <https://doi.org/10.1080/02602938.2019.1639613>

Almalkawi, I., Jester, R., Terry, L., 2018. Exploring mentors' interpretation of terminology and levels of competence when assessing nursing students: An integrative review. Nurse education today, 69, 95–103. <https://doi.org/10.1016/j.nedt.2018.07.003>

Bachmann, L., Groenvik, C.K.U., Hauge, K.W., Julnes, S., 2019. Failing to Fail nursing students among mentors: A confirmatory factor analysis of the Failing to Fail scale. Nursing Open, 6(3), pp. 966-973. <https://doi.org/10.1002/nop2.276>

Boateng, B.A., Bass, L.D., Blaszak, R.T., Farrar, H.C., 2009. The development of a competency-based assessment rubric to measure resident milestones. Journal of graduate medical education, 1(1), pp.45-48. [https://dx.doi.org/10.4300%2F01.01.0008](https://dx.doi.org/10.4300/01.01.0008)

Bork, C.E.,1993. Research in Physical Therapy. Philladelphia: J.B. Lippincott Co

Bowling, A., 2014. Research methods in health: investigating health and health services. UK: McGraw-Hill Education.

Bradshaw, C., Pettigrew, J., Fitzpatrick, M., 2019. Safety first: Factors affecting preceptor midwives experiences of competency assessment failure among midwifery students. Midwifery, 74, pp.29-35. <https://doi.org/10.1016/j.midw.2019.03.012>

Brown, L., Douglas, V., Garrity, J., Shepherd, C.K., 2012. What influences mentors to pass or fail students. Nursing Management, 19(5), 16-21. <https://doi.org/10.7748/nm2012.09.19.5.16.c9260>

Buck, A. J., Gross, M., Hakim, S., Wainblatt, J., 1993. Using the Delphi process to analyse social policy implementation- a post hoc case of from vocational rehabilitation. *Policy Sciences*, 26(4), 271-288. <https://doi.org/10.1007/BF00999473>

Burkhart-Kriesel, C., Trauman, K., Rice, T., 2011. Benchmarking to Build. Sustainable Communities: “Expect, Like and Love to See”. University of Nebraska. Retrieved from: <http://extensionpublications.unl.edu/assets/pdf/ec860.pdf>.

Butler, M.P., Cassidy, I., Quillinan, B., Fahy, A., Bradshaw, C., Tuohy, D., O’Connor, M., Mc Namara, M.C., Egan, G., Tierney, C., 2011. Competency assessment methods–tool and processes: a survey of nurse preceptors in Ireland. Nurse Education in Practice, 11(5), 298-303. <https://doi.org/10.1016/j.nepr.2011.01.006>

Calman, L., Watson, R., Norman, I., Redfern, S. and Murrells, T., 2002. Assessing practice of student nurses: methods, preparation of assessors and student views. *Journal of Advanced Nursing*, *38*(5), pp.516-523. <https://doi.org/10.1046/j.1365-2648.2002.02213.x>

Cassidy, S., Coffey, M., Murphy, F., 2017. ‘Seeking authorization’: a grounded theory exploration of mentors’ experiences of assessing nursing students on the borderline of achievement of competence in clinical practice. Journal of advanced nursing, 73(9), 2167-2178. <https://doi.org/10.1111/jan.13292>

Cassidy, S., Coffey, M., Murphy, F., 2020. Transparency of assessment decision-making when students are not meeting required levels of proficiency in clinical practice. Nurse Education in Practice, 43(2), pp. 102711. <https://doi.org/10.1016/j.nepr.2020.102711>

Cockett, A., Jackson, C., 2018. The use of assessment rubrics to enhance feedback in higher education: An integrative literature review. Nurse education today, 69, 8-13. <https://doi.org/10.1016/j.nedt.2018.06.022>

Cooney, F., Stebbings S.N., Roxburgh, M., Mayo, J., Keen, N., Evans, E., Meehan, T.C., 1995. Integrating Nursing Research and Practice, Part 11 - a Delphi study of nursing practice priorities for research-based solutions. Nursing Practice in New Zealand, 10(1), 22-27.

Croft, H., Gilligan, C., Rasiah. R., Levett-Jones, T., Schneider, J., 2019. Current Trends and Opportunities for Competency Assessment in Pharmacy Education–A Literature Review. Pharmacy, 7(2), PP. 67. <https://doi.org/10.3390/pharmacy7020067>

DeBrew, J.K., Lewallen, L.P. 2014. To pass or to fail? Understanding the factors considered by faculty in the clinical evaluation of nursing students. Nurse Education Today, 34(4), 631-636. <https://doi.org/10.1016/j.nedt.2013.05.014>

De Vos, M., Baartman, L., Van Der Vleuten, C., De Bruijn, E., 2019. Exploring how educators at the workplace inform their judgement of students’ professional performance. Journal of Education and Work, 32:8, 693-706. <https://doi.org/10.1080/13639080.2019.1696953>

Diamond, I.R., Grant, R.C., Feldman, B.M., Pencharz, P.B., Ling, S.C., Moore, A.M., Wales, P.W., 2014. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. Journal of clinical epidemiology, 67(4), pp.401-409. <https://doi.org/10.1016/j.jclinepi.2013.12.002>

Donaldson, J.H., Gray, M., 2012. Systematic review of grading practice: is there evidence of grade inflation? Nurse Education in Practice, 12(2), 101-114. <https://doi.org/10.1016/j.nepr.2011.10.007>

Donohoe, H.M., Needham, R.D., 2009. Moving best practice forward: Delphi characteristics, advantages, potential problems, and solutions. International Journal of Tourism Research, 11(5), 415-437. <https://doi.org/10.1002/jtr.709>

Duffy, K., 2003. Failing students. A qualitative study of factors that influence the decisions regarding assessment of students‟ competence in practice London; NMC, London. Retrieved from: <http://www.nmcuk.org/Documents/Archived%20Publications/1Research%20papers/Kathleen_Duffy_Failing_Students2003.pdf>

Finch, J., 2017. “… It’s Just Very Hard To Fail A Student…”: Decision-Making And Defences Against Anxiety–An Ethnographic And Practice-Near Study Of Practice Assessment Panels. *Journal of Social Work Practice*, *31*(1), 51-65. <https://doi.org/10.1080/02650533.2016.1158156>

Fitzgerald, M., Gibson, F., Gunn, K., 2010. Contemporary issues relating to assessment of pre-registration nursing students in practice. Nurse education in practice, 10(3), 158-163. <https://doi.org/10.1016/j.nepr.2009.06.001>

Frentsos, J.M., 2013. Rubrics role in measuring nursing staff competencies. Journal for nurses in professional development, 29(1), 19-23. <https://doi.org/10.1097/NND.0b013e31827d0a9c>

Gallagher, P., Smith T., Ousey, K., 2012. Problems with competence assessment as it applies to student nurses. Nurse education in practice, 12(6), 301-303. <https://doi.org/10.1016/j.nepr.2012.05.014>

Goodman, C. M., 1987. The Delphi technique: a critique. Journal of advanced nursing, 12(6), 729-734. <https://doi.org/10.1111/j.1365-2648.1987.tb01376.x>

Govaerts, M.J.B., Van de Wiel, M.W.J., Schuwirth, L.W.T., Van der Vleuten, C.P.M., Muijtjens, A.M.M., 2013. Workplace-based assessment: raters’ performance theories and constructs. Advances in Health Sciences Education, 18(3), 375-396. <https://doi.org/10.1007/s10459-012-9376-x>

Gray, M.A., Donaldson, J., 2009. Exploring issues in the use of grading in practice; literature review. Edinburgh Napier University, Edinburgh.

Gray, J.R., Grove, S.K., Sutherland, S., 2017. Burns and Grove's The Practice of Nursing Research: Appraisal, Synthesis, and Generation of Evidence (8th ed.). St. Louis, MO: Elsevier Health Sciences.

Haines, C., Dennick, R., da Silva, J.A.P., 2013. Developing a professional approach to work-based assessments in rheumatology. Best Practice and Research Clinical Rheumatology, 27(2), 123-136. <https://doi.org/10.1016/j.berh.2013.02.006>

Hasson, F., Keeney, S., McKenna, H., 2000. Research guidelines for the Delphi survey technique. Journal of advanced nursing, 32(4), 1008-1015. <https://doi.org/10.1046/j.1365-2648.2000.t01-1-01567.x>

Heaslip, V., Scammell, J.M., 2012. Failing underperforming students: the role of grading in practice assessment. Nurse Education in practice, 12(2), 95-100. <https://doi.org/10.1016/j.nepr.2011.08.003>

Hill, K.Q., Fowles, J., 1975. The methodological worth of the Delphi forecasting technique. Technological Forecasting and Social Change, 7(2), 179-192. [https://doi.org/10.1016/0040-1625(75)90057-8](https://doi.org/10.1016/0040-1625%2875%2990057-8)

Hishiti, E., Amkongo, M., Karera, A., 2020. Radiography students’ experiences of practical evaluations at University of Namibia, The South African Radiographer, 58(1), 20-26.

Holloway, I., Wheeler, S., 1996. Qualitative research for nurses. Oxford: Blackwell Science.

Hunt, L.A., 2014. Failing Securely: Enabling mentors to fail underperforming student nurses in practical assessments. Doctoral dissertation, Birmingham City University.

Hunt, L.A., McGee, P., Gutteridge, R., Hughes, M., 2012. Assessment of student nurses in practice: A comparison of theoretical and practical assessment results in England. Nurse Education Today, 32(4), 351-355. <https://doi.org/10.1016/j.nedt.2011.05.010>

Hughes, L.J., Mitchell, M.L. and Johnston, A.N., 2020. Moving forward: Barriers and enablers to failure to fail–A mixed methods meta-integration. *Nurse Education Today*, p.104666. <https://doi.org/10.1016/j.nedt.2020.104666>

Immonen, K., Oikarainen, A., Tomietto, M., Kääriäinen, M., Tuomikoski, A.M., Kaučič, B.M., Filej, B., Riklikiene, O., Vizcaya-Moreno, M.F., Perez-Cañaveras, R.M. and De Raeve, P., 2019. Assessment of nursing students' competence in clinical practice: a systematic review of reviews. *International journal of nursing studies*, *100*, p.103414. <https://doi.org/10.1016/j.ijnurstu.2019.103414>

Jokelainen, M., Tossavainen, K., Jamookeeah, D., Turnen, H., 2013. Finish and British mentors’ conceptions of facilitating nursing students’ placement learning and professional development. Nurse Education in Practice, 13(1), 61-67. <https://doi.org/10.1016/j.nepr.2012.07.008>

Keeney, S. Hasson, F., MvKenna, H., 2011. The Delphi Technique in Nursing and Health Research. Chichester: Wiley-Blackwell.

Kennedy, S. and Chesser-Smyth, P., 2017. Assessment of undergraduate nursing students from an Irish perspective: Decisions and dilemmas? *Nurse education in practice*, *27*, pp.95-100. <https://doi.org/10.1016/j.nepr.2017.08.017>

Landeta, J., 2006. Current validity of the Delphi method in social sciences. Technological Forecasting and Social Change, 73(5), 467-82. <https://doi.org/10.1016/j.techfore.2005.09.002>

Larocque, S., Luhanga, F.L., 2013. Exploring the issue of failure to fail in a nursing program. International Journal of Nursing Education Scholarship, 10(1), 1-8. <https://doi.org/10.1515/ijnes-2012-0037>

Lemmer, B., 1998. Successive surveys of an expert panel: research in decision‐making with health visitors. Journal of advanced nursing, 27(3), 538-545. <https://doi.org/10.1046/j.1365-2648.1998.00544.x>

Lincoln, Y.S., Guba, E.G.,1985. Naturalistic Inquiry. Newbury Park, CA: Sage Publications.

Linstone, H.A., 1978. The Delphi technique. Handbook of futures research. *J. Fowlers. Westport, CT, Greenwood Press*, pp.273-300.

McCarthy, B., Murphy, S., 2008 Assessing undergraduate nursing students in clinical practice: do preceptors use assessment strategies? Nurse Education Today, 28(3), 301-313. <https://doi.org/10.1016/j.nedt.2007.06.002>

McDowell, L., Wakelin, D., Montgomery, C., King, S., 2011. Does assessment for learning make a difference? The development of a questionnaire to explore the student response. Assessment and Evaluation in Higher Education, 36(7), 749-765. [https://srhe.tandfonline.com/author/Montgomery%2C+Catherine](https://srhe.tandfonline.com/author/Montgomery%2C%2BCatherine)

McKenna, H. P.,1994. The Delphi technique: a worthwhile approach for Nursing? Journal of Advanced Nursing, 19(6), 1221-1225. <https://doi.org/10.1111/j.1365-2648.1994.tb01207.x>

Miller, C., 2010. Improving and enhancing performance in the affective domain of nursing students: Insights from the literature for clinical educators. Contemporary Nurse, 35(1), 2–17. <https://doi.org/10.5172/conu.2010.35.1.002>

Montgomery, K., 2002. Authentic Tasks and Rubrics: going beyond traditional assessments in College Teaching. College teaching, 50(1), 34-40. <https://doi.org/10.1080/87567550209595870>

Neary, M., 2001. Responsive assessment: assessing student nurses’ clinical competence. Nurse Education Today, 21(1), 3-17. <https://doi.org/10.1054/nedt.2000.0508>

Nicola-Richmond, K., Butterworth, B. and Hitch, D., 2017. What factors contribute to failure of fieldwork placement? Perspectives of supervisors and university fieldwork educators. *World Federation of Occupational Therapists Bulletin*, *73*(2), 117-124. <https://doi.org/10.1080/14473828.2016.1149981>

Nugent, O., Lydon, C., Part, S., Dennehy, C., Fenn, H., Keane, L., Prizeman, G. and Timmins, F., 2020. Who is failing who? A survey exploration of the barriers & enablers to accurate decision making when nursing students’ competence is below required standards. *Nurse Education in Practice*, *45*, p.102791. <https://doi.org/10.1016/j.nepr.2020.102791>

Nursing and Midwifery Council (2018) Standards for student supervision and assessment. NMC: London.

Owens, C., Ley, A., Aitken, P., 2008. Do different stakeholder groups shape mental health service priorities? A four arm Delphi Study. Health Expectations, 11(4), 418-431. [https://dx.doi.org/10.1111%2Fj.1369-7625.2008.00492.x](https://dx.doi.org/10.1111/j.1369-7625.2008.00492.x)

Oyinloye, O.M., Imenda, S.N., 2019. The Impact of Assessment for Learning on Learner Performance in Life Science. Eurasia Journal of Mathematics, Science and Technology Education, 15(11), p. em1775. <https://doi.org/10.29333/ejmste/108689>

Schuwirth, L.W., Van der Vleuten, C.P., 2011. Programmatic assessment: from assessment of learning to assessment for learning. Medical teacher, 33(6), 478-485. <https://doi.org/10.3109/0142159X.2011.565828>

Sharma, G., 2017. Pros and cons of different sampling techniques. International Journal of Applied Research, 3(7),749-752.

Shipman, D., Roa, M., Hooten, J., Wang, Z.J., 2012. Using the analytic rubric as an evaluation tool in nursing education: the positive and the negative. Nurse education today, 32(3), 246-249. <https://doi.org/10.1016/j.nedt.2011.04.007>

Skulmoski, G.J., Hartman, F.T. and Krahn, J., 2007. The Delphi method for graduate research. *Journal of Information Technology Education,* 6 (2), 1-20. <https://www.learntechlib.org/p/111405/>

Strauss, C., Taylor, B.L., Gu, J., Kuyken, W., Baer, R., Jones, F., Cavanagh, K., 2016. What is compassion and how can we measure it? A review of definitions and measures. Clinical psychology review, 47, 15-27. <https://doi.org/10.1016/j.cpr.2016.05.004>

Sumison, T., 1998. The Delphi Technique: an adaptive research tool. The British Journal of Occupational Therapy, 61(4), 153-156. [https://doi.org/10.1177%2F030802269806100403](https://doi.org/10.1177/030802269806100403)

Trede, F., Smith, M., 2012. Challenges of assessment in practice-based education. In Practice-Based Education (pp. 187-198). Brill Sense.

Von der Gracht, H. A., 2008. The Delphi technique for futures research. The Future of Logistics: Scenarios for 2025, 21-68. <https://doi.org/10.1007/978-3-8349-9764-7_3>

Watson, R., Stimpson, A., Toping, A., Porock, D., 2002. Clinical competence assessment in nursing: a systematic review of the literature. Journal of advanced nursing, 39(5), 421-431. <https://doi.org/10.1046/j.1365-2648.2002.02307.x>

Wiewiora, A., Kowalkiewicz, A., 2019. The role of authentic assessment in developing authentic leadership identity and competencies. Assessment and Evaluation in Higher Education, 44(3), pp.415-430. <https://doi.org/10.1080/02602938.2018.1516730>

Wiggins, G., 1989. A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70(9), 703-713. [https://doi.org/10.1177%2F003172171109200721](https://doi.org/10.1177/003172171109200721)

Winkelbauer, S., 2020. An Authentic, Practice-Based Assessment as a Catalyst for Continuous Professional Development. Pharmacy, 8, 1-15. <https://doi.org/10.3390/pharmacy8010015>

Wood, M.J., Ross-Kerr, J.C., 2011. Basic steps in planning nursing research: From question to proposal (7th ed.). London: Jones and Bartlett.

Yonge, O., Myrick, F., Ferguson, L., 2011. The process of developing a framework to guide rural nurse preceptors in the evaluation of student performance. Nurse education in practice, 11(2), 76-80. <https://doi.org/10.1016/j.nepr.2011.01.001>