Delphi study to determine the key qualities consultant histopathologists look for in their trainees

Daniel J Brierley [1], Paula M Farthing,[2] Sandra Zijlstra-Shaw[3]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jclinpath-2019-206345).

[1]Unit of Oral and Maxillofacial Pathology, University of Sheffield, Sheffield, UK
[2]University of Sheffield, Sheffield, UK
[3]Unit of Oral and Maxillofacial Pathology, University of Sheffield School of Clinical Dentistry, Sheffield, UK

Correspondence to
Dr Daniel J Brierley, Unit of Oral and Maxillofacial Pathology, University of Sheffield, Sheffield S10 2TA, UK; d.j.brierley@sheffield.ac.uk

Received 21 November 2019
Revised 28 January 2020
Accepted 1 February 2020
Published Online First
10 April 2020

To cite: Brierley DJ, Farthing PM, Zijlstra-Shaw S. *J Clin Pathol* 2020;**73**:642–647.

ABSTRACT

Aims A Delphi study to triangulate and determine the relative importance of the key qualities of trainees identified from qualitative interviews that sought to understand how consultant histopathologists determine diagnostic competences in trainees.

Methods Twelve participants were purposively chosen for the Delphi to form an expert panel of relevant stakeholders. Participants were asked to score and rank the items presented to them.

Results A total of 22 out of 27 of the key qualities of trainees (items) reached 'consensus in' after round 2 suggesting participants were able to agree that the majority of the items identified in the qualitative interviews were important to diagnostic competence. Five items reached 'no consensus'. Participants did not suggest any additional items. Participants particularly valued qualities of reflection and professionalism and trainees who understood the process of reaching a diagnosis and how their pathological report could impact on patient care.

Conclusions This study has triangulated findings from our qualitative interviews and show that consultants value a wide variety of qualities when determining diagnostic competence in their trainees. The judgement is complex and is therefore best assessed longitudinally and on a number of cases, so consultants can look for consistency of both approach to diagnosis and of trainee behaviour.

INTRODUCTION

We have previously discussed the key qualities that consultants look for when determining diagnostic competence in histopathology trainees.[1] This was described as a longitudinal judgement, encompassing the qualities of the process of making a diagnosis (the process) as well as personal characteristics (the person). Expectations of competence are, to some extent, stage dependent and guided by the Royal College of Pathologists (RCPath) guidelines. However, the judgement is complex, incorporating both evidence and feelings about trainees. Diagnostic competence appears to manifest in consultants 'trusting' their trainees to be independent practitioners (figure 1).

In order to triangulate the findings from these qualitative interviews and to quantify the importance of the qualities identified, a Delphi study was undertaken. This methodology was in line with our overall theoretical framework of Social Judgement Theory (SJT). In brief, SJT creates a way to understand clinical judgements and how individual cues (or facets of information) are used in making judgements. It has been used in several clinical fields, including the use of clinical information by nurses and the diagnostic ability of clinicians[2 3]

The Delphi method was originally used in the 1950s by the RAND Air Force Corporation in America.[4] Within the field of medical education, it is the most commonly selected consensus group method, accounting for approximately 75% of papers.[5] Delphi involves using a panel of experts to determine the importance of items. It has several integral features: anonymity, iteration, controlled feedback, statistical group response and structured interaction.[6]

OBJECTIVES

- To determine if the items (key qualities of a trainee) identified from the qualitative interviews are valued by experts.
- To determine the relative importance of individual items.
- To identify any additional items.

METHOD

Participants were purposively selected in order to create an 'expert' panel. It is important to stress that panel members should offer expertise and a range of viewpoints to be most useful.[7] In line with recommendations,[5] it was felt that training programme directors (TPDs) were most knowledgeable about diagnostic competence, and they were approached in addition to other consultants who had roles within the RCPath or the Deanery (other than TPD or educational supervisor. All potential participants received an invitation email and participant information sheet detailing the nature of the study.

There is no stated recommended number of participants for Delphi studies[5]; however, 12 or more is considered reasonable.[4]

The final Delphi panel consisted of 12 participants from both general and oral and maxillofacial pathology (table 1).

The items included in the Delphi study were generated from the qualitative interview data previously described.[1] The themes and subthemes were re-examined to develop descriptions that would form items that could be rated by participants. The Delphi was first piloted with an experienced TPD who was not part of the main study. Through an iterative process, a list of 27 items was created (see online supplementary table l).

Participants were asked to rate the items in terms of importance on a 10-point Likert-type item scale where 10 represented 'definitely important' and 1 represented 'definitely not important'.

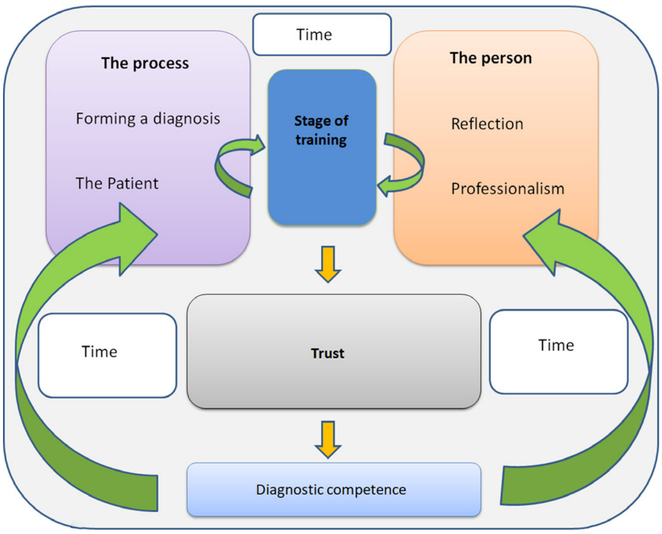

Figure 1 Conceptual framework illustrating how consultants pathologists determine diagnostic competence in their trainees.

Participants were also asked to rank the three most important and three least important items when determining diagnostic competence in trainees.

For both rating and ranking judgements, participants were encouraged to justify their scores or offer other items that were not listed in the Delphi questionnaire. This information was used to help understand the rationale behind the consensus.

Consensus was defined as 'consensus in' when >70% of participants scored the item as 8–10 (definitely important), <15% had scored it 1–3 (definitely not important) and the IQR was 2 or less. 'Consensus out' was when >70% of participants scored an item 1–3, <15% participants scored it 8–10 and the IQR was 2. All other combinations were considered to be 'no consensus', in accordance with previous research.[8 9]

Wilcoxon matched pairs signed rank test was used to determine if there was a significant difference in opinion per item per round. This shows whether the consensus is 'stable'[10] and provides an objective way of determining the number of rounds needed.

All communication between the researcher and Delphi participants occurred over private, individual emails so participants remained anonymous to one another. Participants were requested to return their completed Delphi form within 3 weeks.

For each of the rounds the following was calculated:

- Response rate.
- List of items that reached consensus/non-consensus.
- Median and IQR for each item.
- Top three ranking items and bottom three ranking items.

Table 1 Summary of Delphi panel participants

Delphi panel background	
Gender	Six males; six females
Geography	Panellists were selected from across the UK.
Roles (past or current)	
Training programme director (TPD)	8
Educational supervisor (ES)	11
Additional college or deanery role (other than TPD/ES)	6

In addition, free-text comments were analysed to see if items needed amending or additional items required adding to Delphi.

In round 2, participants were asked to rate and rank the items again and were given information from the previous round including:

- Participant's own scores.
- Panel score.
- Range of scores for that item.
- List of items that had reached consensus/non-consensus.
- Any amendments/additions to original items from round 1.

Basic descriptive statistics and inferential statistics (Wilcoxon matched pairs signed rank test) were calculated, using GraphPad Prism version 7.00 for Windows, GraphPad Software, La Jolla, Californiam USA, www.graphpad.com. Free-text comments were analysed thematically. The response rates for rounds 1 and 2 were 100%.

RESULTS AND DISCUSSION

Five participants provided feedback on the 27 items in round 1, which mostly related to stage of training. For example:

> *"The questions do not differentiate between a year 1 trainee and a year 5 one. Some of my answers might be different if stage of training was included.*
> *Many of your statements need clarification with the year of training – because this will make a huge difference".*

The following items (under their respective headings) were therefore amended for round 2 (bold text indicates amendment):

Forming a diagnosis

Item 9: the diagnosis **is commensurate with stage of training.**

Item 10: trainee consistently produces **a diagnosis commensurate with stage of training.**

Trust

Item 19: trainee can be trusted to carry out macroscopic examination and 'cut-up' independently **commensurate with stage of training.**

Item 21: **Senior trainees** can be trusted to report independently (the consultant does not check the report and the trainee authorises it).

It was decided not to give a stage of training in the first round of the Delphi to see if participants would identify this issue for themselves. It is therefore useful to see that the Delphi panellists identified stage of training as integral to the judgement process, as it was also integral to our model of diagnostic competence (figure 1).

A further comment from a participant was:

> *"Exhibiting ownership of cases, i.e. acting to take responsibility and not as a consultant's assistant pathologist, independent initiative to sort cases out".*

Therefore, item 20 under 'Trust' was amended to reflect this stance:

Item 20: trainee can be trusted to manage cases, **showing ownership and initiative** (however, the consultant will check the reports and authorise them).

The median, IQR and the items that reached consensus 'in' or 'no consensus' for rounds 1 and 2 are shown in table 2. table 3 gives details of which items reached consensus 'in' or 'no consensus' in rounds 1 and 2.

A total of 17 out of the 27 items reached 'consensus in' after round 1, and this increased to 22 after round 2. Ten items reached 'no consensus' after round 1 and this reduced to 5 after

Table 2 Comparison of median, IQR and which items reached consensus in rounds 1 and 2 under consensus

Item	Median round 1	Median round 2	IQR round 1	IQR round 2	Consensus round 1	Consensus round 2
1	9.5	9	3	1.25	🟥	🟩
2	7	7	3.25	1.5	🟥	🟥
3	8	8	1.25	1	🟩	🟩
4	8	6.5	3	2	🟥	🟥
5	9.5	9	2.25	0.25	🟥	🟩
6	10	9	2	1.5	🟩	🟩
7	10	9	2	2	🟩	🟩
8	10	9	1	1	🟩	🟩
9	7	8	4.25	1.25	🟥	🟥
10	8	8	3.25	1	🟥	🟥
11	9	8.5	1.25	1.25	🟩	🟩
12	9	9	1	0.25	🟩	🟩
13	9	9	1	0.25	🟩	🟩
14	9	9	0.5	1	🟩	🟩
15	8	8	1	1.25	🟥	🟩
16	10	10	1	1.25	🟩	🟩
17	10	9.5	1.25	1.25	🟩	🟩
18	10	9.5	1	1	🟩	🟩
19	7.5	8	2	1.25	🟥	🟩
20	8	8	1	1	🟩	🟩
21	9	8.5	2.25	1	🟥	🟩
22	10	10	0.25	1	🟩	🟩
23	5.5	6	2.25	1.25	🟥	🟥
24	9.5	9	1.25	0.25	🟩	🟩
25	10	9	2	1.25	🟩	🟩
26	10	9.5	1	1	🟩	🟩
27	9.5	9	2	1	🟩	🟩

Light grey row indicates 'consensus in' and dark grey 'no consensus'.

round 2. No items reached 'consensus out' in either round. In addition, the IQR stayed the same or reduced for 23 of the 27 items in round 2 indicating that participants showed increasing agreement in round 2, and therefore, a further five items reached 'consensus in'.

Participants did not suggest any additional items to the original 27 items provided, suggesting the qualitative interviews explored the topic of diagnostic competence to the correct depth. Wilcoxon matched pairs signed rank test showed no significant difference between round 1 and round 2 responses for each of the twenty-seven items ($p>0.05$), so further rounds were not conducted.

Tables 4 and 5 show items with the highest and lowest median score and rank respectively at the end of the Delphi.

The results indicate that the majority of items identified in the qualitative interviews were also valued by the Delphi panel with 22 of the 27 items reaching 'consensus in' after two rounds. This helps to triangulate and validate our previous findings. Only five items failed to reach 'consensus in', and of those five, none were deemed unimportant enough to be scored 'consensus out'.

It can be seen from table 3 that all the items under 'The patient's report', 'Reflection' and 'Professionalism' reached 'consensus in' suggesting that these categories were particularly important to participants. Linked to this, those items that had the highest median score or rank (table 4) were mostly within the category 'Reflection'. Consultants appear to value highly trainees who know their own limitations, do not act beyond their competency and know when to ask for help. Our findings are akin to the Cambridge model of competency, which illustrates how personal relationships affect performance including aspects of professionalism and reflection.[11]

Table 4 also shows that items 16 'Trainee is aware of limitations' and 22 'Trainee asks for help/second opinion when needed' received both high median scores and ranks. which demonstrates how personal qualities are highly valued. This is supported by the inclusion of items 17 'Trainee reflects on their own limitations/performance' and 18 'Trainee shows evidence of improvement following reflection' in the top three ranked or rated items. Similarly, the highly ranked or rated items 13 'Trainee follows sound pathological principles to reach a diagnosis', and 26 'Report ensures the clinician/surgeon receives the appropriate message' indicate that the diagnostic process itself and its outcome are also highly valued. Nonetheless, a total 16 items had median scores of 9 or above (table 2) suggesting there were many items considered 'definitely important'. Indeed, one participant commented '*My least important selection is a bit arbitrary as all are important*'. Given that 89% of items had median scores of eight or more, this comment appears justified. Many of the items in this study also had significant overlap and impact on one another, which is reflected in a Delphi participant commenting that '*Some of these are complexly interrelated and whilst on their own are not important, are functions of other traits*'. Therefore, scoring and ranking items, in some ways, oversimplifies how complex the judgement task is. It has been recognised that implementing outcome-based medical education can be challenging in terms of translating the complexities of medical practice into meaningful assessment strategies and curricula.[12] These complexities are clearly reflected in the

Table 3 Details of individual items reaching consensus 'in' or 'no consensus' in rounds 1 and 2

Item	Consensus round 1	Consensus round 2
Stage of training		
1. The stage of training for example, ST1 versus ST2.	🟥	🟩
2. RCPath curricula and training guidelines.	🟥	🟥
3. Placements/cases assessed longitudinally over a period of time.	🟩	🟩
4. Individual cases in an examination setting.	🟥	🟥
Professionalism		
5. Trainee is organised and timely when conducting themselves in the department, for example, does not lose cases, is aware of turnaround times, triages urgent cases.	🟥	🟩
6. Trainee communicates with all staff appropriately, for example, effectively, timely and politely.	🟩	🟩
7. Trainee works as part of a team, for example, works effectively with colleagues and does not create conflict.	🟩	🟩
8. Trainee is motivated and has a good attitude.	🟩	🟩
Forming a diagnosis		
9. The diagnosis is correct.	🟥	🟥
10. Trainee consistently produces a correct diagnosis.	🟥	🟥
11. The histopathology report is accurate and does not contain factual errors or omit important information.	🟩	🟩
12. Trainee shows evidence of clinico-pathological correlation.	🟩	🟩
13. Trainee follows sound pathological principles to reach a diagnosis.	🟩	🟩
14. Trainee has appropriate level of basic knowledge.	🟩	🟩
15. Trainee has appropriate level of up-to-date knowledge, for example, latest papers/datasets.	🟥	🟩
Reflection		
16. Trainee is aware of their own limitations.	🟩	🟩
17. Trainee reflects on their own limitations/performance.	🟩	🟩
18. Trainee shows evidence of improvement following reflection.	🟩	🟩
Trust		
19. Trainee can be trusted to carry out macroscopic examination and 'cut-up' independently.	🟥	🟩
20. Trainee can be trusted to report cases (however, the consultant will check the reports and authorise them).	🟩	🟩
21. Trainee can be trusted to report independently (the consultant does not check the report and the trainee authorises it).	🟥	🟩
22. Trainee asks for help/second opinion when needed.	🟩	🟩
23. Your opinion of a trainee's diagnostic competence is related to an overall impression you have of them.	🟥	🟥
24. Your opinion of a trainee's diagnostic competence is related to bringing all the evidence together and triangulating findings from exams, workplace-based assessments and day-to-day work.	🟩	🟩
The patient's report		
25. Report is useful to the clinician/surgeon, for example, it does not contain unnecessary information or detail.	🟩	🟩
26. Report ensures the clinician/surgeon receives the appropriate message.	🟩	🟩
27. Report is organised appropriately and well-written.	🟩	🟩

Consensus 'in' is indicated in light grey and 'no consensus' is indicated in dark grey
RCPath, Royal College of Pathologists.

Table 4 Comparison of highest scoring and ranked items after round 2

Item	Score	Item	Rank
Reflection 16. Trainee is aware of their own limitations.	10	**Reflection** 16. Trainee is aware of their own limitations/	1st
Trust 22. Trainee asks for help/second opinion when needed	10		
Reflection		**Forming a diagnosis** 13. Trainee follows sound pathological principles to reach a diagnosis.	2nd
17. Trainee reflects on their own limitations/performance.	9.5	**Trust** 22. Trainee asks for help/second opinion when needed.	3rd
18. Trainee shows evidence of improvement following reflection.	9.5		
The patient's report 26. Report ensures the clinician/surgeon receives the appropriate message.	9.5		

Note: some items have been shortened for reasons of clarity.

Table 5 Comparison of lowest scoring and ranked items after round 2

Item	Score	Item	Rank
Trust 23. Your opinion of a trainee's diagnostic competence is related to an overall impression.	6	**Trust** 23. Your opinion of a trainee's diagnostic competence is related to an overall impression.	=1st
Timing 4. Individual cases in an examination setting.	6.5	**Timing** 4. Individual cases in an examination setting.	=1st
Stage of training 2. RCPath curricula and training guidelines.	7	**Stage of training** 2. RCPath curricula and training guidelines.	3rd

Note: some items have been shortened for reasons of clarity.
RCPath, Royal College of Pathologists.

data supporting our conceptual model of diagnostic competence, where the various factors involved are all linked to one another. By separating these into competencies, the complexities of performance are unlikely to be captured.

As mentioned previously, 'following sound pathological principles' was ranked highly (item 13) (table 4), reflecting that the understanding of cases and the underpinning elements that support a diagnosis are of great importance. This is supported by a participant comment '*If a trainee completes with sound principles in box 13 then their training has been successful in my opinion*'.

In contrast, getting 'the diagnosis correct commensurate with the stage of training' (item 9, round 2 median score 8) or even consistently correct (item 10, round 2 median score 8) (table 2) did not reach 'consensus in'. Even with the amendment to items 9 and 10 after round 1 to incorporate 'commensurate with the stage of training', these items still did not reach 'consensus in'. This further highlights that the approach to diagnosis and understanding how the diagnosis is reached are actually more important than the end product: the diagnosis. Furthermore, aspects of reflection and professionalism appear to impact on this process.

'Stage of training', 'Timing' and 'Trust' each had one item that did not reach 'consensus in' and 'Forming a diagnosis' had two (table 3). Given that TPDs were represented in both the interviews and the Delphi, and no new items were suggested; the simplest explanation for this is that some of the items within these categories were not considered as important as other items.

Items 2 'RCPath curricula and training guidelines', 4 'Individual cases in an examination setting' and 23 'Your opinion of a trainee's diagnostic competence is related to an overall impression' were the lowest scoring and ranked items at the end of the Delphi and also failed to reach consensus 'in' or 'out'. The qualitative interviews identified that curricula and training guidelines were used in part to make judgements on trainee diagnostic competence (item 2). Even though it was one of the lowest ranked items, item 2 still appeared relatively important because it had a median score of 7 after round 2. The panel members were all considered 'experts' so their extra experience may mean they are very familiar with training pathways and do not need to refer to guidelines or curricula as frequently as others. Whether these views represent the wider view (eg, consultants who are less experienced) is unclear.

Regarding item 23, participants were rather equivocal with regard to whether 'feelings' or 'impressions' were important when judging trainee diagnostic competence (median score 6). It is possible that their experience(s) made them unconsciously able to measure competence and they did not recognise this ability or its importance. In contrast to item 23, item 24, which suggested trainee diagnostic competence is 'related to bringing all the evidence together and triangulating findings from exams, workplace-based assessments and day-to-day work', had a median score of 9. It therefore appears that assimilating the evidence is more important than 'feelings' even though 'feelings' do sometimes exist.

Item 4 'individual cases in an examination setting' had a low score/rank, and this is not surprising as our qualitative data suggested competence should be assessed longitudinally rather than on a single episode. In support of this item 3, 'Placements/cases assessed longitudinally over a period of time' did reach consensus in. Participants valued repeated demonstrations of 'process' and 'person' to develop trust. This echoes the sentiments of Oerlemans and colleagues, who found that clinical supervisors appreciate consistent behaviours when assessment of a trainee is based on a series of observations.[13] This is because the judgement ecology is not consistent, where stage of training, environment, attitudes and emotions can all affect the outcome.[14 15] In line with the literature, diagnostic competence involves the ability to manage ambiguous problems, tolerate uncertainty and make decisions with limited information.[16] However, it is important to stress that examinations still have a role in training. While they may not be good at measuring qualities such as professionalism or reflection, they can provide an external, quality assured assessment to determine if trainees are able to apply their knowledge in their field of practice.

The importance of trusting trainees appears to be an extremely important aspect of diagnostic competence given that items 19 'Trainee can be trusted to carry out macroscopic examination and "cut-up" independently', 20 'Trainee can be trusted to report cases (however, the consultant will check the reports and authorise them)' and 21 'Trainee can be trusted to report independently (the consultant does not check the report and the trainee authorises it)' all reached consensus. Consultants appear to want to review 'all the evidence' and spend time with trainees before they feel completely happy for them to report independently. The 'time' spent with a trainee to assess competence and delegation of tasks is in line with the work of Dijksterhuis and colleagues,[17] who concluded that the depth of acquaintance with a trainee is the most important factor affecting when to delegate work in postgraduate training. One Delphi study identified 25 facets of competence valued by physician educators when entrusting tasks to trainees.[18] They concluded that their findings were 'useful for the development of a valid method for assessing medical graduates' readiness for clinical practice'. This mirrors other work that has studied the motivation behind entrustment decisions and suggested research is required to identify tools that enable faculty to justify their entrustment decisions.[19] Indeed, entrustable professional activities (EPAs) are being used increasingly in specialty training where there is a focus on what tasks consultants feel happy entrusting to trainees. This acts as an indicator of their development and ability to operate in the workplace, rather than looking individual competencies.[20] Our research suggests a role for EPAs in the assessment of diagnostic competence, but the practicalities of determining what and when certain activities can be delegated to trainees requires further work. There is also the wider issue of working with the current 'risk adverse' culture within the medical profession and creating detailed guidelines to help inform exactly how independent reporting and similar forms of delegation can be put into practice safely. For example, despite two studies citing the possible positive contribution that increased responsibility can bring, these studies also concluded that trainees are currently rarely exposed to it.[21 22]

LIMITATIONS

In reality, there are infinite types of trainee and context to consider and scores and ranks of the individual items might not reflect this complexity in the workplace. The definition of consensus that was used in this study is not an absolute, as it is a subjective benchmark. It is for the regulatory bodies and assessors to determine 'how important' something has to be before it is included in any assessment strategies.

CONCLUSION

This study has triangulated findings from our qualitative interviews. No new items were suggested by participants, suggesting the qualitative interviews explored diagnostic competence in sufficient depth. Consideration should be given to incorporating these qualities into assessment tools used in histopathology, such as evidence of reflection, which was highly valued. In addition, these findings suggest the assessment of competence in histopathology is best viewed longitudinally and on a number of cases, rather than 'snap-shots' captured on workplace-based assessments. Diagnostic competence culminates in consultants trusting their trainees to perform certain tasks independently. Curricula should focus on what trainees do in the workplace rather than demonstration of individual competencies. Further work is needed to determine the pedagogic approach and feasibility of delivering these findings within assessments.

Take home messages

- ► This Delphi study triangulates our previous qualitative research and suggests our model of diagnostic competence should be used in training.
- ► Diagnostic competence should be assessed longitudinally rather than on individual cases.
- ► Assessment tools should place more emphasis on reflection and professionalism as these qualities are highly valued by consultants when determining competence in their trainees.
- ► Consideration should be given to how delegation of work and independent reporting can be used to monitor trainee development.

This paper stems from a PhD thesis, which is published online at http://etheses.whiterose.ac.uk/.

Handling editor Dhirendra Govender.

Contributors DJB: literature review, study design, data analysis, discussion and conclusions. PMF: study design, discussion, conclusions and proof-reading. SZ-S: study design, discussion, conclusions and proof-reading.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. No further data are available.

ORCID iD
Daniel J Brierley http://orcid.org/0000-0003-0152-2372

REFERENCES

1 Brierley DJ, Farthing PM, Zijlstra-Shaw S. How consultants determine diagnostic competence in histopathology trainees. *J Clin Pathol* 2019;72:622–9.
2 Thompson CA, Foster A, Cole I, *et al*. Using social judgement theory to model nurses' use of clinical information in critical care education. *Nurse Educ Today* 2005;25:68–77.
3 Wigton RS, Hoellerich VL, Patil KD. How physicians use clinical information in diagnosing pulmonary embolism: an application of conjoint analysis. *Med Decis Making* 1986;6:2–11.
4 Murphy MK, Black NA, Lamping DL, *et al*. Consensus development methods, and their use in clinical Guideline development. *Health Technol Assess* 1998;2:i–iv.
5 Humphrey-Murto S, Varpio L, Gonsalves C, *et al*. Using consensus group methods such as Delphi and nominal group in medical education research. *Med Teach* 2017;39:14–19.
6 Jones J, Hunter D. Qualitative research: consensus methods for medical and health services research. *BMJ* 1995;311:376–80.
7 Bloor M, Sampson H, Baker S, *et al*. Useful but no Oracle: reflections on the use of a Delphi group in a multi-methods policy research study. *Qualitative Research* 2015;15:57–70.
8 Tsichlaki A, O'Brien K, Johal A, *et al*. Development of a core outcome set for orthodontic trials using a mixed-methods approach: protocol for a multicentre study. *Trials* 2017;18:366.
9 Harman NL, Bruce IA, Kirkham JJ, *et al*. The importance of integration of stakeholder views in core outcome set development: otitis media with effusion in children with cleft palate. *PLoS One* 2015;10:e0129514.
10 Holey EA, Feeley JL, Dixon J, *et al*. An exploration of the use of simple statistics to measure consensus and stability in Delphi studies. *BMC Med Res Methodol* 2007;7:52.
11 Rethans J-J, Norcini JJ, Barón-Maldonado M, *et al*. The relationship between competence and performance: implications for assessing practice performance. *Med Educ* 2002;36:901–9.
12 Frank JR, Snell LS, Cate OT, *et al*. Competency-Based medical education: theory to practice. *Med Teach* 2010;32:638–45.
13 Oerlemans M, Dielissen P, Timmerman A, *et al*. Should we assess clinical performance in single patient encounters or consistent behaviors of clinical performance over a series of encounters? A qualitative exploration of narrative trainee profiles. *Med Teach* 2017;39:300–7.
14 Essers G, van Dulmen S, van Weel C, *et al*. Identifying context factors explaining physician's low performance in communication assessment: an explorative study in general practice. *BMC Fam Pract* 2011;12:1–8.
15 Ginsburg S, Bernabeo E, Ross KM, *et al*. "It depends": results of a qualitative study investigating how practicing internists approach professional dilemmas. *Acad Med* 2012;87:1685–93.
16 Schon DA. *The reflective practitioner*. New York: Basic Books, 1983.
17 Dijksterhuis MGK, Voorhuis M, Teunissen PW, *et al*. Assessment of competence and progressive independence in postgraduate clinical training. *Med Educ* 2009;43:1156–65.
18 Wijnen-Meijer M, van der Schaaf M, Nillesen K, *et al*. Essential facets of competence that enable trust in graduates: a Delphi study among physician educators in the Netherlands. *J Grad Med Educ* 2013;5:46–53.
19 Sterkenburg A, Barach P, Kalkman C, *et al*. When do supervising physicians decide to entrust residents with unsupervised tasks? *Acad Med* 2010;85:1408–17.
20 ten Cate O, Chen HC, Hoff RG, *et al*. Curriculum development for the workplace using entrustable professional activities (EPAs): AMEE guide No. 99. *Med Teach* 2015;37:983–1002.
21 Davey DD, Talkington S, Kannan V, *et al*. Cytopathology and the pathology resident: a survey of residency program directors. *Arch Pathol Lab Med* 1996;120:101–4.
22 Pascal RR. Graded responsibility of residents in anatomic pathology. A survey and commentary. *Am J Clin Pathol* 1993;100:S41–3.