PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Supervised Learning Events in the Foundation Programme: A UK-wide narrative interview study
AUTHORS	Rees, Charlotte; Cleland, Jennifer; Dennis, Ashley; Kelly, Narcie; Mattick, Karen; Monrouxe, Lynn

VERSION 1 - REVIEW

REVIEWER	Jennifer Weller University of Auckland New Zealand
REVIEW RETURNED	16-Jul-2014

GENERAL COMMENTS	Thank you for the opportunity to review this manuscript. The authors have clearly gone to a great deal of trouble to get to explore the experiences of and reactions to SLEs and WBPAs of trainees and trainers. I found the data interesting and it does add something to the understanding of how these assessments work for early medical graduates. It wasn't clear what message to take away from this other than confusion reigns. Table Legends don't always contain explanations of all abbreviated terms. I did struggle with the description of the methodology and the structure of the results. I could be accused of being a "methodological purist" perhaps. In particular, the "quantitative" (e.g. "trainers were more likely than trainees to…" conclusions about differences between trainers and trainees and between SLEs and WBPAs seemed difficult to support from the data provided. While the data identified problems, the proposed recommendations don't appear to be based on examples of what works from the data provided. Also, these are the responses of 110 participants and can't be claimed to be representative of all trainees and trainers. Suggest modifying the conclusions. Introduction I don't work in the UK. I don't really know how an SLE is run for Foundation doctors. It would be helpful to explain it. This may go some way to clear up the confusion in Table 2 – while it appears the tools drawn upon by SLE and WBPA overlap, it's unclear why there is a difference in answers to the question "Had experience with SLE tools / WBPA tools" when they are the same tools. I understand this is an open list – perhaps some assumed that answering the question once about experience with DOPS or MiniCEX was enough and they didn't need to write it down again? Please clarify. Methods The reference to participants, by sharing stories, "construct identities and trainee-trainer relationships" promises a very interesting

	analysis. It was disappointing that this involved only a single narrative about a trainee and a DOPS on inserting an IV cannula. Perhaps just change this to reflect a single narrative analysis. Sampling was from Year 1 and 2 of the programme and from different regions and specialities – this may not qualify as "maximum-variation sampling", which would could conceivable aim to identify trainees and trainers who may have extreme views on the issue. Sample size – why 110 participants? Why focus groups for some and interviews for others? Were the interviews mainly for the trainers? Were the focus groups mixed? Data collection The authors describe "narrative interviewing" but the data presented doesn't seem to refer to specific experiences – can the authors clarify how the reported data was "grounded in actual lived experiences". Data analysis Please clarify what is meant by "Thematic Framework Analysis". My understanding is that The Framework Method is a form of thematic analysis, but this doesn't seem to be what the authors have actually done here. It seems more that the "Framework" has been determined by the different research questions. In the abstract, the methodology is described as "a qualitative and quantitative thematic and discourse analysis and narrative analysis etc." This is all pretty confusing. None of the four different methods of analysis is explained in sufficient detail to understand what was done. Please clarify and provide references. I wonder if, rather than trying to present the analysis as a single coherent analysis with an overarching methodology, the different approaches could be described as separate sections and the results reported accordingly. Results The first paragraph outlines the major headings for the results and to some extent repeats the research questions and is also partly repeated at the beginning of the different results sections. The language is inconsistent each time it's written, including in the headings for the tables (e.g. "What are participants understandings of SLEs/ WBAs etc" as a heading alternative with "conceptualisation of SLE / WBPA). I suggest it would be less confusing if the same thing was consistently referred to with the same words. I'd also suggest (as indicated above) that these are not presented as seven themes identified by a "thematic framework analysis" but results of different methods of data analysis. RQ1 is really just answers to the first question from what appears in the table. It's difficult to see how the authors can claim that "While SLEs were conceptualised as learning and assessment, WPBAs were typically understood as assessment. Trainers were more likely than trainees..etc". The data presented suggests "many" didn't know what SLE's/WBPAs were, "many" thought they were the same, and "others" thought SLEs may be more formative. We don't know what trainees thought about SLEs as a safety net – perhaps it just didn't come up in the interview. Suggest modifying the claims or providing the data to support these quantitative comparisons. The single example of a trainee being anxious about a WBPA didn't strike me as "striking". I wonder if the authors could provide more data to support their interpretation. RQ2 Please explain what is meant by "fragmentary themes". Table 3 – Please explain in the methods section how the narrative

	were scored as positive or negative. RQ3 – I enjoyed this example of discourse analysis (it would help to have described the method in detail prior, and also the particular choice of this narrative). I was left wondering however if this was a general theme or if this was just this particular narrative. It would seem the claims of surviving, them and us, etc. are interesting but the reader will be hard pressed to know what to make of this single episode.

REVIEWER	Colin Mitchell Dept of Elderly Medicine Imperial College NHS Healthcare Trust UK
REVIEW RETURNED	07-Aug-2014

GENERAL COMMENTS	This is a valuable exploration of attitudes and understanding relating to SLEs in the UK NHS. It has practical implications and adds some depth and rigour to our understanding of how these tools are used in real practice. The conclusions drawn are wide-ranging and it seems this work raises more questions than it resolves, which is not necessarily a bad thing, although I think it's fair to say that it would require a wider sampling of opinions to fully elucidate all the themes developed. However I think there is enough material here and a representative sample of trainers and trainees from which to draw the conclusions made. The tools and processes involved are inextricably linked and both highly contextual which perhaps limits some of the generalisability of this work. The suggestions for practical implementation to address some of the issues raised are workable and reasonable, and all worthy of further study in themselves. I'm please to see this type of work getting a broader audience.

VERSION 1 – AUTHOR RESPONSE

R1: Clarification of criteria for positive, negative narratives required for Table 3. Narratives typically contain numerous elements including ‘evaluation’ (which is the narrator’s commentary on their experience: Labov & Waletsky 1967). We did not use a scoring system or criteria for coding ‘evaluation’. Instead, we developed a set of codes for evaluation inductively as part of our thematic analysis e.g. positive evaluation, negative evaluation, neutral evaluation, unclear and so on. As per the interpretive approach, the analysts coded whole narratives to these codes as they saw appropriate, and depending on what participants said and how they said it. For example, narratives including mostly negative emotional talk (e.g. “it was quite alarming”) would be coded to ‘negative evaluation’ and narratives including mostly positive emotional talk (e.g. “it’s nice to have nice things said about you”) would be coded to ‘positive evaluation’ etc. We have added a brief explanation of this in the results (see ‘the context of SLE and WPBA narratives’ section for RQ2) and the footnote for Table 3.

R1: It wasn’t clear what message to take away from this other than confusion reigns. We are not quite sure how to respond to this comment, as it is general/non-specific. We think the key messages for the paper are clear in the abstract and the first section of the discussion. For example, confusion exists about what SLEs are; SLEs and WPBAs are conducted in diverse ways; differences exist in terms of trainees' and trainers' experiences; participants construct their identities and relationships as they make sense of their experiences; and participants suggested various ways that SLEs could be improved.

R1: Table Legends don't always contain explanations of all abbreviated terms. Thank you. We now include full definitions of abbreviations across all Tables as part of the title and footnotes (e.g. DOPS, Mini-CEX, CBD, SLE, WPBA). See all tables.

R1: I did struggle with the description of the methodology and the structure of the results. I could be accused of being a "methodological purist" perhaps. In particular, the "quantitative" (e.g. "trainers were more likely than trainees to…" conclusions about differences between trainers and trainees and between SLEs and WBPAs seemed difficult to support from the data provided. As we already explain in the 'methodological strengths and challenges' section of the discussion, our study employs a process-orientated qualitative approach and only uses numbers to illustrate patterns in our large qualitative data set of personal incident narratives (n=333); patterns that would be invisible without the use of numbers. We think the quantification of the contextual codes for SLE vs. WPBA narratives (e.g. the evaluation of narratives: see Table 3) does provide clear data illustrating differences between SLEs and WPBAs. In our original submission of the paper, we illustrate these patterns using descriptive statistics only (i.e. frequencies and percentages). However, in our revised paper we go one step further and use univariate statistics (i.e. chi-squared tests) to illustrate that this association is statistically significant. For example, a chi-squared test shows that there is a significant association between evaluation (positive/negative) and type of experience (SLE/WPBA: X2=5.344, df=1, p=.021). However, the patterns identified in the original paper between evaluation (positive/negative) and participant (trainee/trainer) were found to be non-significant. We have added these additional statistics to the revised paper (see 'the context of SLE and WPBA narratives' section of the results) and have revised our interpretations accordingly.

R1: While the data identified problems, the proposed recommendations don't appear to be based on examples of what works from the data provided. The proposed recommendations are primarily based on the participants' responses to the question: how do you think that SLEs could/should be improved? (RQ4) Data in response to this question is of course linked to participants' problematic experiences and how such problematic experiences can be rectified. However, the suggestions are also based on good experiences shared by trainees and trainers within the interviews. We now make this clearer in our implications section: "Our recommendations are based on key findings from our research (both what works and what does not work) and comments from our clinical reference group (see acknowledgements)".
Also, these are the responses of 110 participants and can't be claimed to be representative of all trainees and trainers. Suggest modifying the conclusions. We agree. Indeed, at no point do we make any claims that our sample of participants is 'representative' of all trainees and trainers in the UK. Qualitative research seeks to elucidate participants' views and experiences and thus embraces subjectivities and multiple realities. We have attempted to identify breadth of views and experiences through maximum-variation (not representative) sampling. Qualitative research, however, sometimes makes claims for transferability (not generalizability). Given that our findings are similar from one site to the next, we have made some claims for transferability in the discussion section. We already state that our sample is not reflective of GP and nurse trainers and so caution is needed extrapolating our findings beyond hospital-based medicine. However, we have now added another comment stating that our sample is not necessarily representative of all UK trainers and trainees.

R1: I don't work in the UK. I don't really know how an SLE is run for Foundation doctors. It would be helpful to explain it. This may go some way to clear up the confusion in Table 2 – while it appears the tools drawn upon by SLE and WBPA overlap, it's unclear why there is a difference in answers to the question "Had experience with SLE tools / WBPA tools" when they are the same tools. I understand this is an open list – perhaps some assumed that answering the question once about experience with DOPS or MiniCEX was enough and they didn't need to write it down again? Please clarify. We have added a touch more explanatory detail to our introduction that we hope reduces any confusion: "Trainees are encouraged to complete a minimum number of SLEs evenly spread throughout their placements with different trainers and covering diverse acute and long-term clinical problems [1]". We also give page numbers for the Collins report to indicate clearly where the reader should look for full details. Furthermore, we re-word slightly some headers in Table 1 e.g. from "had experience with SLEs tools" to "had experiences with tools as SLEs", to make it clearer that the tools are the SAME for both SLEs and WPBAs but are supposed to be used differently (i.e. formatively in SLEs). We hope this is now adequately clarified in our revised manuscript.

R1: The reference to participants, by sharing stories, "construct identities and trainee-trainer relationships" promises a very interesting analysis. It was disappointing that this involved only a single narrative about a trainee and a DOPS on inserting an IV cannula. Perhaps just change this to reflect a single narrative analysis. Thank you for the reviewer for this comment: this is an issue that we struggled with ahead of submitting our paper to BMJOpen. Narrative analysis is typically lengthy so given the word limits for the journal we knew we had space only to include one narrative and its in-depth analysis. Note that we have published several papers recently, which follow the same format (including only one in-depth narrative analysis e.g. Rees et al. 2013, Rees et al. 2014). What we decided to do in our revision is to add a paragraph of text ahead of our introduction of this one narrative that draws on multiple narratives across the data, and includes further fragmentary quotes from various narratives, to illustrate that key issues around discourse (e.g. pronominal, emotional and metaphoric talk) are relevant across our data, and so are not just specific to this one narrative. Furthermore, we have added a reference to our end-of-award report (Rees et al. 2013), which includes another in-depth narrative analysis – so interested readers are directed to that report should they wish to read another in-depth narrative analysis of a trainer narrative.

R1: Sampling was from Year 1 and 2 of the programme and from different regions and specialities – this may not qualify as "maximum-variation sampling", which would could conceivable aim to identify trainees and trainers who may have extreme views on the issue. We politely disagree with the reviewer on this point. We went to great efforts to include a diverse sample by participant type, participant demographics, year of training (trainees), level of teaching experience (trainers), specialties, and geography, and so we think we do have a maximum-variation sample. It would have been impossible (nor desirable) to identify and recruit trainees/trainers across three regions in the UK based on their actual views and experiences of SLEs/WPBAs, because we would not know what their views and experiences were until we interviewed them (so post-recruitment). All one can do as a qualitative researcher is to maximize variation in your sample in terms of other variables (such as all those mentioned above) with the assumption that it should lead to variation in views and experiences.

R1: Sample size – why 110 participants? Why focus groups for some and interviews for others? Were the interviews mainly for the trainers? Were the focus groups mixed? Why 110 participants: When we wrote our grant application for funding to the AOMRC, we stated that we aimed to conduct 27 focus groups, each involving approximately 4 participants, across three sites (England, Wales and Scotland) resulting in an overall study sample of n=108 (with 36 participants per participant type). Although adequate sample sizes for qualitative research vary from one to more than one hundred depending on research aims and methods, some qualitative scholars advocate for around 30 interview participants as a minimum (e.g. Adler & Adler 2012), so this planned sample size exceeded this minimum number. Given our experience conducting focus group and interview studies within the context of medical education, we thought that this sample size was the maximum that we could possibly include given the financial and time constraints of the grant, a key limitation to sample pools as discussed by other qualitative scholars (Adler & Adler 2012). We ended up conducting more interviews and fewer focus groups than expected (because of the difficulties in getting groups of extremely busy people together at the same time), so we exceeded slightly our original target, giving a final sample of 110. We have added a statement to this affect in our sampling and recruitment section: “We interviewed 110 participants (34 F1s, 36 F2s, and 40 trainers: see Table 1 for participants’ characteristics). This overall sample and sub-samples far exceeded the minimum sample size of 30 advocated by some qualitative scholars (e.g. Adler & Adler 2012). Furthermore, we considered this to be the maximum number of participants we could feasibly interview given the time and financial constraints of our grant, another pragmatic consideration discussed by qualitative researchers (Adler & Adler 2012).

Why focus groups and interviews: We have added a statement addressing this in our revised study design section: “We employed focus groups wherever possible because they can lead to richer data due to group dynamics (e.g. synergism) but individual interviews were also utilised because of the difficulties in getting groups of clinicians together”.

Were the interviews mainly for trainers: We conducted interviews with trainers and trainees, but the reviewer is assuming correctly: the majority of trainers took part in an individual interview rather than a focus group, because we found it nearly impossible to organize focus groups with multiple trainers due to pressures on their time. We have added this detail to the first line of our data collection section: “34 individual and 3 group interviews with trainers; 21 individual and 16 group interviews with trainees”.

Were the focus groups mixed: The vast majority of the focus groups were homogenous in terms of type of participant (i.e. they either included trainers or trainees from year-specific groups). We conducted only 2 mixed groups and these were both trainee groups including FY1 and FY2 trainees). Any analyses looking at the differences between FY1s and FY2s excluded these two mixed groups. We have added the following statement to the revised data collection section: “Note that all focus groups bar two were homogenous in terms of type of study participant (i.e. trainer or year-specific trainee groups)”.

R1: The authors describe “narrative interviewing” but the data presented doesn’t seem to refer to specific experiences – can the authors clarify how the reported data was “grounded in actual lived experiences”. We used narrative interviewing techniques to elicit stories of experience (SLE and WPBA experiences) from trainers and trainees. With the exception of two key questions (what’s your understanding of SLEs/WPBAs? And what’s your suggestions for improving SLEs?), all data were ‘narrative’ and therefore data is grounded in participants’ lived experiences rather than eliciting participants’ generalized opinions and perceptions about SLEs/WPBAs. Granted, we only present one full narrative in the paper because of word count limitations (see further comments in relation to this issue below), but all of the data for RQ2 comes from these lived experiences. We have made this clearer in the results section for RQ2 by adding the comment: “… all pertaining to participants’ lived experiences of SLEs/WPBAs”.

R1: Please clarify what is meant by “Thematic Framework Analysis”. My understanding is that The Framework Method is a form of thematic analysis, but this doesn’t seem to be what the authors have actually done here. It seems more that the “Framework” has been determined by the different research questions. In the abstract, the methodology is described as “a qualitative and quantitative thematic and discourse analysis and narrative analysis etc.” This is all pretty confusing.
None of the four different methods of analysis is explained in sufficient detail to understand what was done. Please clarify and provide references. We are sorry that the reviewer was confused by our multi-analysis approach. We agree that this was confusing in our original submission as we had tried to keep this section brief in order to keep the paper within the word limits set by the journal. We have revised our data analysis section by adding more explanatory detail about each type of analysis, and we now include a reference for all types of analysis (i.e. Framework analysis, discourse analysis and

narrative analysis). We also include one of our own references where we have used a similar multi-analysis approach before (Monrouxe & Rees 2012), so that interested readers can read more about the different types of analyses and their combination. We hope this section is now clearer and more aligned with the abstract.

R1: I wonder if, rather than trying to present the analysis as a single coherent analysis with an overarching methodology, the different approaches could be described as separate sections and the results reported accordingly. We politely disagree with the reviewer on this point. We think it is much more appropriate (and clearer to the reader) to present the findings of our study by research question rather than by the type of analysis. There is no clear relationship between the RQ and the type of analysis (e.g. Framework analysis was used in relation to all four research questions, narrative analysis was used in relation to one of the four research questions) and so on. Therefore, to present the findings by type of analysis would lead, we think, to a messy articulation of the results that don't map neatly onto the research questions. We hope that by addressing the reviewer's previous point about the data analysis section, the paper will be much clearer.

R1: The first paragraph outlines the major headings for the results and to some extent repeats the research questions and is also partly repeated at the beginning of the different results sections. The language is inconsistent each time it's written, including in the headings for the tables (e.g. "What are participants understandings of SLEs/ WBAs etc" as a heading alternative with "conceptualization of SLE / WBPA). I suggest it would be less confusing if the same thing was consistently referred to with the same words. We purposely use headings aligned with our research questions in the results section because we think this will make the study findings by research question very clear to the reader. We agree that some of the inconsistencies in our language could be confusing so we now use the term "understandings" (instead of "conceptualizations") throughout the revised paper.

R1: I'd also suggest (as indicated above) that these are not presented as seven themes identified by a "thematic framework analysis" but results of different methods of data analysis. See above response to this issue.

R1: RQ1 is really just answers to the first question from what appears in the table. It's difficult to see how the authors can claim that "While SLEs were conceptualised as learning and assessment, WPBAs were typically understood as assessment. Trainers were more likely than trainees..etc". The data presented suggests "many" didn't know what SLE's/WBPAs were, "many" thought they were the same, and "others" thought SLEs may be more formative. We don't know what trainees thought about SLEs as a safety net – perhaps it just didn't come up in the interview. Suggest modifying the claims or providing the data to support these quantitative comparisons. The single example of a trainee being anxious about a WBPA didn't strike me as "striking". I wonder if the authors could provide more data to support their interpretation. The findings presented for RQ1 are the findings for the first question asked in the interview but the headings in Table 2 do reflect the sub-themes that we identified as researchers (through the Framework analysis). We completely understand what the reviewer here is saying about our use of the word "many" etc. in relation to these sub-themes. While we think it entirely appropriate to quantify participants' narratives for RQ2, we do not think it appropriate to quantify (using numbers) the definitions volunteered by our participants about SLEs and WPBAs (even though we have this data from our Atlas-Ti coding), because such quantification in relation to questions that may or may not have been repeatedly asked to different participants in an interview or focus group setting may be somewhat misleading. Therefore, what we have done in this first results section of the paper, is to change our language so that it is more 'hedged' and thus modifying our claims e.g. we change "While participants demonstrated a range of understandings for SLEs" to "While participants volunteered a range of understandings for SLEs" and we change "another striking difference…" to "another apparent difference we identified…".

R1: Please explain what is meant by "fragmentary themes". We now add the following explanation in parenthesis: "(i.e. themes that cross-cut all narratives)".

R1: Table 3 – Please explain in the methods section how the narrative were scored as positive or negative. See earlier point. We now explain this in the results section and as a footnote to Table 3.

R1: RQ3 – I enjoyed this example of discourse analysis (it would help to have described the method in detail prior, and also the particular choice of this narrative). I was left wondering however if this was a general theme or if this was just this particular narrative. It would seem the claims of surviving, them and us etc. are interesting but the reader will be hard pressed to know what to make of this single episode. We have added an extra sentence about discourse analysis and a couple of references in our data analysis section. We now explain why we selected this particular narrative for inclusion in the paper in the results section. The issues discussed as part of this narrative (e.g. pronominal, metaphoric, and emotional talk) are not just relevant to this specific narrative but are relevant to other narratives across the data, and we now include further examples by way of illustrative quotes ahead of this one example. Also, the reader is sign-posted to our end-of-award report should they wish to read another in-depth narrative analysis.

R2: The conclusions drawn are wide-ranging and it seems this work raises more questions than it resolves, which is not necessarily a bad thing, although I think it's fair to say that it would require a wider sampling of opinions to fully elucidate all the themes developed. We agree with the reviewer to a degree. We have added another recommendation to our revised paper in terms of further research along these lines: “Further interview research is required using wider sampling (e.g. capturing GP experiences) to more fully elucidate the themes identified in this paper”.

R2: The tools and processes involved are inextricably linked and both highly contextual which perhaps limits some of the generalisability of this work. As we have already responded, at no time do we make claims about the generalizability of our findings, although we do make qualitative claims in relation to transferability. As stated above, we have added another comment to our methodological challenges section stating that our sample is not necessarily representative of all UK trainers and trainees.

VERSION 2 – REVIEW

REVIEWER	Jennifer Weller University of Auckland New Zealand
REVIEW RETURNED	20-Sep-2014

GENERAL COMMENTS	The added details on methodolgy and results clarify the manuscript - an interesting read.