Proposed Frontline Evaluation Analysis Plan for Simulated Practice Data

19th October, 2015

The Frontline evaluation research team at CASCADE, Cardiff University
(contact Jonathan Scourfield scourfield@cardiff.ac.uk)

The full protocol for the DfE-commissioned independent evaluation of the Frontline pilot was published online in September 2014, with an updated revised version published in January 2015. The protocol is available from the CASCADE website (http://sites.cardiff.ac.uk/cascade/publications-2/cascade-publications/ - CASCADE Paper 01). To be as transparent as possible, the research team at Cardiff University have devised the following analysis plan for the evaluation of the simulated practice data. To begin, we introduce this element of the evaluation, repeating some of the information in the protocol.

Objective 3 of the evaluation (as set by DfE) is to 'measure objectively how well Frontline prepares participants to be outstanding social workers'. For this element of the evaluation a quasi-experimental study has been set up. The research team have built on the use of simulated practice in social work education by Marian Bogo and colleagues (2014) in Toronto, Canada. This is a model for the standardised assessment of social work practice that mirrors the Objective Structured Clinical Examination (OSCE) used in medical education. Frontline trainees are being compared with social work students who are about to qualify on regular programmes. Two comparison groups are used. These are (1) students in high undergraduate (UG) tariff universities and (2) a sample of students from a range of other mainstream programmes (both UG and postgraduate [PG]). The first group have been highlighted not because of any assumption of programme quality but because of necessary assumptions about the postgraduate student market, with the aim of identifying students most similar to Frontline trainees in terms of academic background. There are no data available via the Higher Education Statistics Agency (HESA) to inform the selection of comparator universities; HESA do not systematically collect data from postgraduates on A-level grades or class of first degree. However, if the postgraduate student market is at all similar to the UG market, we might assume that the best qualified students will be drawn to Masters programmes in the same universities that have the highest entry standards at UG level, meaning that these students would be the most comparable with Frontline participants in terms of academic background. Frontline's admissions criteria include an upper second degree or higher and at least 300 UCAS points in top three A-levels or equivalent. The Guardian newspaper university league table (The Guardian, 2013) was consulted. Thirteen universities which teach PG social work were in the bracket of 400+ points for all-subject UG entry tariff. In identifying universities to approach about participation in the evaluation, all-subject tariff was considered more relevant than the tariff for UG social work specifically, as it is the former that affects position in league tables.

Criteria were agreed by consensus for assessing qualifying students' skills via simulated service user interviews and written reflections on these. A 'Delphi' process was undertaken with equally-weighted groups of social work academics, practice educators, practitioners and service users, in order to reach agreement about a system for scoring practice quality. The Delphi method consists of a series of individual consultations with domain experts, interspersed with controlled feedback of the experts' opinions (Dalkey and Helmer,1963). The academics were recruited via advertisement to the Joint Universities Council Social

Work Education Committee (JUC SWEC) email list. Although similar advertisements were put out for practitioners (via the College of Social Work) and practice educators (via the National Organisation of Practice Teachers), adverts did not generate sufficient interest so practitioners and practice educators involved with the Cardiff MASW programme were recruited. Service users were recruited via the user-led organisation for care-experienced young people Voices from Care. All of these participants had experience of social workers when they were looked after by the local authority and several had also been involved with children's services as parents.

The Delphi group considered the assessment tools for generic social work skills that have been developed and validated by Marian Bogo and colleagues in Canada (http://research.socialwork.utoronto.ca/hubpage/resources-2), slightly slimmed down to reduce the burden on assessors. The Bogo *et al*. criteria were mapped on to the Professional Capabilities Framework, the Health and Care Professions Council's standards of proficiency and the Chief Social Worker's Knowledge and Skills document and found to be compatible, albeit they are only concerned with the range of capabilities which can be assessed via a simulated interview and written reflection and they do not cover the full range of tasks encompassed by these frameworks. The Delphi group agreed in the first round of consultation that the Bogo *et al*. criteria were acceptable for assessing qualifying social workers in the UK. A few minor edits were made to the language in the Bogo *et al*. criteria to ensure their translation to a UK context.

Evaluation participants will be assessed on the basis of their performance in two simulated interviews with actors - one a parent scenario and one a teenage child - and written reflections on these interviews. Performance in the simulated interviews is judged on the basis of ten separate criteria, each scored from one to five (lowest to highest). The assessment of the written reflections is scored in a similar way using six separate criteria. The criteria are listed in Table 1 overleaf. As noted above, the simulated practice assessment does not cover all the qualities required for social work practice. It does not, for example, assess someone's ability to function effectively within an organisation and it does not assess social scientific knowledge in depth. Rating of audio recordings will be done by a pool of practice assessors, with two assessors for each recording and all assessors rating a selection of recordings from each of the three groups, but with no knowledge of group membership.

Table 1: Simulated practice assessment criteria*, as agreed by Delphi process

Practice assessment
The student develops and uses a collaborative relationship
- Introduction
- Response to service user: general content and process
- Response to service user: specific to situation
- Focus of interview
The student conducts an assessment of the person in their environment
- Presenting problem
- Systemic assessment
- Strengths
The student sets the stage for collaborative goal setting
The student demonstrates cultural competence
Overall assessment of the simulated interview
Written reflection assessment
Student is able to conceptualise their practice/make use of knowledge
Content: How students theoretically conceptualise substantive issues in the scenario and for their practice
Content: How students conceptualise issues of culture and diversity in their practice
Process: How students' past knowledge and experience impact their approach to the case
Student is able to assess their own practice
Cognitive: what students focus on and talk about regarding their performance
Student is able to think about their professional development
Learning: What students focus on and talk about regarding their learning
Growth: What students say about how they would integrate this experience into their practice

***** Taken from the work of Bogo *et al*. http://research.socialwork.utoronto.ca/hubpage/resources-2

Each participant in the simulated practice was also asked to complete a questionnaire covering, amongst other things, previous experiences of simulated practice; any pressures experienced during the social work programme; and previous academic achievements. All simulated practice tests have now been completed. Table 2 below contains the final numbers participating and response rates. Data analysis is due to take place in November 2015.

Table 2: Recruitment of simulated practice participants

Group	*n* eligible	*n* completed	Response rate
Higher tariff universities (n=6) PG	121	36	30%
Other universities (n=5) UG+PG	173	30 (13 PG/17 UG)	17%
Frontline	103	49	48%
Total	397	115	29%

The numbers of social work students participating are lower than was originally envisaged, as the aim was for 70 in each group. This is not necessarily a cause for concern, as the power of statistical tests does not decrease linearly with sample size. Representativeness is important to

assess, however, and the research team are currently investigating this, with aggregate data requested from all programmes on the final grades of students participating in the simulated practice compared with their whole cohort. On the basis of the numbers in table 2, we can evaluate the statistical power of the analysis.

Basic analysis plan

In the evaluation we are primarily interested in three way comparisons between Frontline participants, PG students from high tariff universities and UG+PG students from other universities. If we assume that all three groups have the same standard deviation in test scores, we can use analysis of variance (ANOVA) to compare the means across all three groups at once. Use of ANOVA to estimate statistical power is in keeping with the original sample size calculation for the study protocol, although as noted below there is a case for using non-parametric statistical tests. The power calculation can refer to the pilot which was conducted with 25 Masters students: 16 first years and 9 second years. Using the standard deviation of scores from the pilot as our estimate (SD=0.37), we would require a difference of at least 0.21 (on a 5-point scale) in mean practice quality scores between the Frontline group and any one of the two non-Frontline groups to find a statistically significant result. For the written reflection (SD=0.46), it should be possible to detect a difference of at least 0.26 on a 5-point scale. This is the statistical power to detect a small to medium effect size of 0.4 (Cohen's d). Another way of putting it is to imagine a score of 0-100, as for an academic assignment. In this scenario score differences of 5% or larger in interview assessment and 7% or larger in written reflection assessment should be statistically significant (i.e. unlikely to occur by chance if there was no real difference in scores in the population of Frontline and non-Frontline students).

These are worst case scenarios in terms of statistical power[1]; under other scenarios it should be possible to detect even smaller differences between groups. The non-parametric equivalent of the ANOVA (i.e. Kruskal-Wallis test) which we propose to use in the actual evaluation analysis has similar or better statistical power than the ANOVA under most situations whilst retaining distinct advantages (Hecke 2012). In summary, we are fairly confident that given the sample size the analysis will be able to detect meaningful differences in scores in the actual evaluation.

The Kruskal–Wallis[2] test will be used to evaluate the likelihood that any differences in scores between the three groups could have occurred purely as a result of random variation or measurement error. To conduct multiple tests of statistical significance would not be desirable, because of the risk of false positives. Therefore the Kruskal-Wallis test will only be used for comparison of total scores for practice quality and written reflections. Additional differences between the three groups in any of the 16 individual scoring criteria (mentioned above) will be presented as descriptive statistics only.

[1] To clarify, if the high tariff university group had a mean score equal to the grand mean score of the sample then in this scenario we would expect ANOVA to have the least power in picking up differences between the Frontline group and any non-Frontline group.

[2] There is no guarantee that the distribution of the total scores within each group will be normally distributed. Since we do not know beforehand what the distribution of total scores will be it is advisable to use a non-parametric test like the Kruskal-Wallis test instead of a parametric test like ANOVA or the t test.

Further analysis issues

The simulated practice is intended to assess the impact of the Frontline programme compared to other social work programmes. However, prior to their studies, Frontline participants differed from other social work students in many ways. For instance, all eligible applicants to the Frontline programme had to have at least 300 UCAS points for their top three A-levels, and HESA data on UGs and our own questionnaire data for PGs show that only a minority of students on regular social work programmes achieved this tariff. In short, selection into the Frontline programme was not random. It is difficult under these circumstances to evaluate whether any differences in performance in the Frontline evaluation are due to the Frontline programme itself or due to any selection effects.

Different starting points could have implications for like-for-like comparison between the groups. In order to assess whether there are comparable individuals across the different programmes, we will attempt to create a matched sample of Frontline participants and individuals from other programmes. The participants in the sample will be matched based on either GCSE English/Maths grades or A-level tariffs, as well as their undergraduate degree classification. Unmatched cases will not be used in the analysis. It will also be possible to construct comparative samples matched on reported pressures affecting the experience of social work education, such as caring responsibilities and external employment. Both results using matched samples and the full sample will be reported. It is doubtful that matching will solve the problem of selection effects entirely.

There were eight actors used in the simulated interviews and, due to unforeseen circumstances, three of the actors were unavailable to participate in simulated interviews with the Frontline participants. This raises issues if evaluation scores using these three actors are systematically different from scores using the other actors. It is doubtful that this will be the case but simple bivariate tests looking at the scores of non-Frontline participants who had these actors and those that did not can help us decide.

References

Bogo, M., Rawlings, M., Katz, E. and Logie, C. (2014) *Using Simulation in Assessment and Teaching. OSCE Adapted for Social Work.* Alexandria, VA, Council on Social Work Education.

Dalkey, N. and Helmer, O. (1963) An experimental application of the Delphi method to the use of experts. *Management Science*, 9, 3: 458-467

Guardian, The (2013) University league table 2014 http://www.theguardian.com/education/table/2013/jun/03/university-league-table-2014

Hecke, T. V. (2012). Power study of Anova versus Kruskal-Wallis test. *Journal of Statistics and Management Systems, 15* (2-3), 241-247.