Insightful Practice: a reliable measure for medical revalidation

APPENDICES (web supplement files)

APPENDIX 1

METHODS

Mapping exercise

The content validity of each type of feedback was examined by participants completing a mapping exercise of each feedback tool's ability to test a doctor's alignment with GMC required attributes. Participating GPs were asked to rate perceived ability of feedback tools (n=5) to test the 12 attributes of Good Medical Practice using a 7-point Likert scale.[5,33] The mapping exercise was completed at the outset and on completion of the study to see if perceptions changed with experience. Descriptive statistics, ANOVA with associated Post–Hoc tests and Generalisability G- theory[9] were used to assess GP agreement and coverage of desired attributes by tools, with changes in perceptions at the beginning and end of the study examined using paired t-tests.

RESULTS

Mapping exercise

Mean GP scores in the mapping exercise (1-7) for each GMC attribute (row) and tool (column) are given in table 3 with a score of 4 as the neutral point on the Likert scale. Inter-rater reliability (participant agreement) was extremely high at 0.99, both before and after TIPP participation. For each GMC attribute (row) tested, there were significant differences identified in participants' assessment on the ability of different tools to test each attribute (P=0.001). Post-hoc tests (Tukey, Tukey's-b) were used to investigate where these significant differences between feedback tools were for each GMC attribute.

MSF was the tool most expected by participants to be the best test for 11/12 attributes. Patient satisfaction questionnaires were perceived to test communication, patient

partnership and respect best. Practice data were expected to test systems to protect patients and improve care. Conversely, knowledge testing was not expected to test any attribute well pre-study but, after experience of it in the study, it was thought to test application of knowledge, experience and the maintaining of professional performance. The suite of feedback was perceived as testing the required spectrum of GMC attributes with at least one tool rated above the neutral point of four for every attribute.

Paired t tests comparing pre- and post-study scores only showed significant differences for patient questionnaires (mean difference 0.17, 95%CI 0.03 to 0.30, t=2.78, 11df, p=0.02) and knowledge tests (mean difference 0.28, 95%CI 0.15 to 0.41, t=4.7, 11df, p=0.001), with participants valuing both more highly in the light of experience of using them.

APPENDIX 2

METHODS

Reliability of Assessment

Descriptive statistics were calculated using SPSS. Reliabilities (internal consistency and inter-rater) of anonymous assessor decisions for AIP (Questions 1-3, 4 and 5) were assessed using Generalisability G- theory and GENOVA.[9,34] 95% confidence intervals for reliabilities were calculated using Fisher's Z_R transformation.[9] Reliability is denoted by a number between zero and one and indicates the proportion of the variance in scores that can be attributed to true differences, as opposed to measurement error. Internal consistency refers to the extent to which the items within the instrument provide consistent information. Inter-rater reliability indicates the extent to which one rater's assessments are predictive of another rater's assessments. Decision (D) studies were conducted to determine the number of assessors required to achieve a reliability of 0.8, as required in high-stakes assessment.[9]

RESULTS (Table 4)

Reliability of Assessment

Internal consistency and inter-rater reliabilities (AIP questions: 1-3) and inter-rater reliability and associated confidence intervals (AIP questions: 4 and 5) were calculated for assessors' judgement on participants' portfolios for a specified number of assessors. Anonymous assessment of satisfactory portfolio performance was highly reliable (G>0.8, suitable for high-stakes assessment)[9] using a 1-7 Likert scale given only two assessors (AIP Q1-3: elements of *insightful practice*) and three assessors (AIP Q4: global rating of *insightful practice*). Dichotomous judgement of the suitability of GPs for revalidation was also highly reliable with four assessors (AIP Q5). Portfolio marking using AIP required a range of 15-45 minutes for each portfolio for five of the six assessors, with the sixth requiring longer than 90 minutes per portfolio.