APPENDIX E

A COMPARISON OF BORDERLINE GROUPS, CONTRASTING GROUPS AND BORDERLINE REGRESSION

As described in the main text, the three main 'judgements on test takers' retrospective standard setting methods are Borderline Groups (BG), Contrasting Groups (CG), and Borderline Regression (BR) (details of these methods are provided in the Glossary, Appendix C). However, it is difficult to find comparisons between all three of these methods employed simultaneously on the same data set. Generally, where comparisons are drawn, they are between two of these methods only.

There is an interesting theoretical question relating to the Borderline Regression method. In most published versions, the Groups are distributed along the abscissa at even intervals. However, there is no obvious reason to assume that this should be the case, and any deviation from even distribution will obviously change the slope of the regression line, and potentially the intercept with the Group vertical. Perhaps a more appropriate distribution along the abscissa might be calculated in logits, or in some manner proportional to the numbers within the population in each Grade, but there is an element of circularity in this argument.

Since we (the research team!) do not have a clear theoretical perspective on this matter, we decided on an empirical exploration of the relationships between these three methods in a particular context for which we had all the necessary data. This was the data for undergraduate medical student OSCEs at Durham University. A total of 31 OSCEs over five administrations were considered (2 years of data for Year 1 OSCEs, with 5 stations each, and 3 years of data for Year 2 OSCEs with 7 stations each). Cut scores were retrospectively calculated by each method (in practice, the Contrasting Groups method had been the one employed).

After careful consideration of how best to present these results, the differences in the percentage cut scores for each combination at each station were calculated. The results are shown below. In summary, the cut scores are lower for BG than either CG or BR methods, with some exceptions. CG and BR cut scores are very similar, with CG being slightly lower. The means and standard deviations of the differences are:

CG-BG mean +2.68, SD 3.34

CG-BG mean -0.96, SD 4.66

BR-BG mean +3.65, SD 5.18

These differences are highly significant between groups by ANOVA ($p>0.00$).

Wood et al (2006) had previously compared Borderline Regression with a Borderline Group approach and found that the overall cut score was lower and the overall pass rate was higher with the Borderline Regression method. This suggests that the relationship between cut scores generated by these different methods may be context specific. As described in the main report, comparisons of Borderline Regression with prospective judgements on test takers generally suggest that cut scores are lower and pass rates higher with Borderline Regression. In no cases that we are aware of, is convincing evidence for the validity for one approach over another presented, and hence we recommend in the main report that there is no compelling case for selecting one such method over another. However, there is an argument for consistency. The only way in which

evidence for retrospective validity for any one standard setting method can be gathered is if the same method is employed over a long period of time. We therefore conclude that there is no compelling case for changing from Borderline Groups to Borderline Regression (a) without modelling the effect of such a change and (b) without determining that the consequences of the change in cut scores and pass rates correspond to increased validity of the outcomes.