Abstract
While Situational Judgement Tests are often presented as addressing fairness issues in assessment there is relatively little work on differential items functioning or differential validity. This study evaluates the use of an SJT in a high stakes context (selection for professional training) and in particular investigates differential group performance at both the item and test level.

Introduction and Objectives
Situational judgement tests have received much research attention recently (e.g. Weekly and Ployhart, 2006). One of the main claims of the method is that it leads to smaller group differences and while these have been studied at the test level there has been less work looking at item level differences using dif techniques. There have been more validity studies published but much of the work is based on student or incumbent populations rather than applicants and there have been few studies looking at differential validity. (McDaniel et al, 2001, Chan & Schmidt, 2005). This paper looks at differential item and validity effects for a situational judgement tests used in the screening of applicants for GP training.

Measures
The shortlisting of doctors for GP training posts uses two machine marked tests: a multiple choice test of clinical problem solving (CPS) and a situational judgement test (SJT) which targets three competencies: (i) Coping with Pressure, (ii) Professional Integrity and (iii) Empathy. SJT questions consist of scenarios typical of GP work and were developed by a group of 20 SMEs, supported by three psychologists. Development of the SJT was described in detail in Patterson et al (2007). The operational test had 50 items to complete within 90 minutes and uses two response formats: rank 4-5 options and choose best 2/3 options.

Shortlisted applicants attend an Assessment Centre (AC) designed to measure the same three competencies and some additional ones. Candidates complete a number of tailored exercises including interactions with simulated patients. The overall AC score is used as the validation criterion. It is based on the sum of assessor ratings across 3 exercises and 5 competencies.
Sample
Just over 8399 candidates completed the tests in February 2007. 4134 (49%) were trained outside the UK and 52% were female. 33% described their ethnic origin as 'White UK or Ireland' with 47% of Asian origin. Because ethnic origin and place of training are confounded in this data, only place of training is considered. 6029 candidates were offered a place at the Assessment Centre. Similar numbers of candidates were assessed in February 2008 but these data have not yet been fully analysed.

Analytic Approach

Differential Item Functioning
Differential item functioning occurs when members of one group have more difficulty answering a test item than equally able members of another group. While this difference is not in itself bias, if it is caused by a factor which is extraneous to what is being measured then it may well be. Good practice in test development identifies items where there is dif and subjects them to further review to better understand the source of the effect. Identifying and understanding dif can be seen as part of test validation (Zumbo, 1999).

Because the scoring of SJT items used in the GP tests results in scores which range from 0 to 20, parametric analyses can be used to evaluate differential item functioning. This is accomplished using a regression approach. Any incremental prediction of item score for group membership, or the interaction of group membership with test score, over test score is a dif finding. Current results suggest that a handful of items within the 50 item test show some dif with respect to both gender and place of training. Results for all items for both place of training and gender will be presented.

Differential Validity
Just as at the item level, differences in difficulty are not of themselves evidence of bias, so at the test level score differences may reflect real differences in capability between groups. Cleary’s (1968) definition that a test is biased only when the same score predicts a different outcome for members of different groups is now well accepted. This is equivalent to showing that the test is valid for all groups, not just the overall sample. This can be tested using a similar approach to that used for dif. The predictive power of the test score in a regression equation is compared to that when group membership is entered into the equation. To identify both intercept and slope effects the interaction term (test score * group membership) is also used. Current results on a restricted sample suggest that the SJT is valid for all groups. Results with additional criterion data will be presented.

Implications
As well as the specific findings with respect the GP SJT, this study will add to our knowledge of the SJT approach and its fairness.

References