Evaluation of three short-listing methodologies for selection into post-graduate training: the case of General Practice in the UK

Patterson F, Baron H, Carr V, Plint S & Lane P.


Corresponding author; Professor Fiona Patterson; f.patterson@city.ac.uk

Abstract

Objective. To evaluate the predictive validity of three short-listing methodologies for use in selecting trainees into post-graduate training in UK General Practice.

Design. This was an exploratory study to compare three short-listing methodologies. Two methodologies were already used for selection purposes including; a clinical problem-solving test and structured application form questions. The third methodology, a new situational judgement test was evaluated alongside the live selection process.

Setting. A predictive validity evaluation was conducted on a sample of 463 participants who were applying for a training post in General Practice in the UK. Participants completed all three assessments and attended a selection centre (SC) at final stage selection.

Main outcome measures. Participant scores on each shortlisting methodology were compared to scores at the SC, comprising work-related simulations.

Results. Results indicate the structured application form questions, clinical problem-solving test and situational judgement test were all valid short-listing methodologies. The situational judgement test was the most effective independent predictor. Both the structured application form questions and situational judgement test add incremental validity over and above the use of the clinical problem solving test alone. Results show the optimum combination for predictive validity is the use of a clinical problem-solving and situational judgement test.

Conclusion. A combination of the clinical problem-solving and situational judgement test is the most efficient battery of instruments for short-listing as unlike the application form questions, these tests are machine-marked. Future research should explore links with work-based assessment once trainees are in post address predictive validity in the long term.
Introduction

Unprecedented changes in post-graduate medical training in the UK, as a result of Modernising Medical Careers\(^1\), have placed more emphasis on selecting doctors into speciality training \(^2,3\). Almost all large scale recruitment uses a multi-stage process to progressively reduce the applicant pool \(^4\). This paper compares the efficiency and effectiveness of three methods used to short-list applicants for selection into postgraduate training. We present this data to encourage debate within the wider medical and scientific community, and to develop a future research agenda for selection methodology.

As with any selection methodology, various psychometric and legal criteria must be satisfied including standardisation, reliability, validity and fairness \(^3,5\). The method must discriminate fairly between candidates and ensure few ‘false negative’ decisions (i.e. rejecting candidates potentially successful at interview). Processes must be cost efficient (especially with large numbers of candidates) where recruiter time and administrative efficiency are important. The methodology should be acceptable to applicants so that procedural fairness is maintained \(^5\). This paper evaluates the validity of methodologies at short-listing stage.

Recruitment into General Practice in the UK

This study evaluates three short-listing methodologies for selection into postgraduate training in General Practice in the UK. The selection system is designed to process several thousand applicants per year \(^6,7\) and the methodology comprises 3 stages; Stage 1, ‘long-listing’ eligibility checks; Stage 2, short-listing via (i) a clinical problem solving test (CPS; a machine marked test developed from an existing item bank\(^1\) in which a candidate applies clinical knowledge to solve a problem reflecting a diagnostic process or developing a management strategy for a patient) and (ii) structured application form questions (AFQ; comprising open-ended questions targeting domains in the person specification, such as empathy, integrity \(^8\), to which candidates were asked to provide a word-limited response). Once short-listed, candidates attend Stage 3, a selection centre using multiple stations including interactive simulations and written exercises. This methodology has been validated however, the process is relatively resource intensive, especially at short-listing with large volume recruitment.

Typically, 5-10\% of applicants are rejected at long-listing with a further 20-30\% selected out at short-listing. This paper reports on an evaluation of a new short-listing methodology, a situational judgement test (SJT); where applicants are presented with written depictions of scenarios they may encounter at work and asked to identify an appropriate response from a list of alternatives. The evaluation considered whether this new approach could improve efficiency by

\(^1\) The original item bank was developed in North Western region in the UK.
replacing the hand marked AFQ (taking 30 minutes of assessor time per candidate) with a machine-marked measure, using a previously agreed scoring frame.

Recent reviews indicate the predictive and incremental validity of SJTs to be well-established \(^9\)\(^-\)\(^1\(^3\). Although the SJT methodology has been validated for use in medical school admissions \(^1\(^4\), to the authors’ knowledge, this is the first application in post-graduate specialty selection. Here, the SJT focused on three non-clinical selection criteria - empathy, integrity and coping with pressure. Figure 1 provides two example items, illustrating different response formats (ranking and multiple best answer).

***INSERT FIGURE 1***

Table 1 provides a summary of the content, the scoring process and related properties for each of the three short-listing methodologies. All are written tests completed independently and under invigilated conditions. By comparing the predictive validity and resource requirements of each, the purpose was to identify the most efficient and effective methodology for future use.

***INSERT TABLE 1 HERE***

As the CPS and AFQ were currently used in selection, the SJT was piloted alongside. Specifically, the following questions are addressed.

1. The CPS is efficient to use but what validity does it offer compared with other short-listing methodologies?
2. The AFQs are relatively time consuming to score requiring trained assessors. Does it add incremental validity beyond that of the CPS?
3. The SJT is a newly designed methodology in this context. Is it valid and does it have incremental validity over the other measures?
4. Which combination of the three measures is the most effective short-listing methodology?

**Method**

**Participants**

Data was collected during the 2006 selection for GP training in the UK. At short-listing stage, candidates completed one of four pilot forms of the SJT together with their operational assessments of the CPS and AFQ. Candidates were fully consented to participate in completing the SJT as it was for evaluation purposes only, and scores were not used in selection decisions. The SJT was completed in advance of the CPS and AFQ.

**Design and Procedure**

Applicants successful at short-listing were invited to a selection centre (SC) on the basis of which job offers were made. Performance at the SC was used as an outcome measure, using the mean scores across three simulation exercises. The regression for SC results on scores at short-
listing from the CPS, AFQ and SJT was examined to establish predictive validity. Hierarchical regression analysis was used to explore incremental validity.

**Psychometric Properties of the Instruments**

The reliability of each assessment methodology was evaluated using Cronbach’s alpha coefficient and the Spearman Brown formula where appropriate. The CPS and AFQ show satisfactory levels of reliability, (CPS $\alpha = .89$; AFQ $\alpha = .78$). Because the SJT was a pilot rather than an operational test, some items were below the standard required for an operational test and were not included in the bank. Results showed 71% of items to be of sufficient psychometric quality to include in an operational test. The Spearman Brown formula was used to estimate the reliability of an operational length test from the items of acceptable quality in each version, ranging from .80 to .83. In preparing SJT items, ten experienced item writers from the examiner panel for Membership of the Royal College of General Practitioners (MRCGP) in the UK, with no previous involvement in the SJT development process, each responded to two forms of the test to provide a five person concordance sample for the response key. This concordance analysis was undertaken to ensure experts were in agreement over the keyed response to each trial item. To indicate concordance between experts Kendall’s W was computed for each ranking item, which showed 85% had a concordance over 0.6 and 71% above 0.7. This indicates substantial and significant levels of agreement between experts.

**Sample**

A total of 524 participants attended the SC, short-listed via the CPS and AFQs. Of these full data was available for 463 candidates. Of the 463 applicants, 178 of respondents were male and 275 were female. 39% described themselves as White (UK and Ireland), 46% Asian, 4.6% Black and 10.4% from other ethnic groups (European, Arab etc). Approximately 58% of the sample was aged 30 or under; 37% were between 31-40 years and 5% over 40 years.

**Results**

**Predictive Validity**

***INSERT TABLE 2 HERE***

All three short-listing methodologies and the SC show score distributions close to normal with effective levels of variability (see Table 2). There were significant correlations between scores on each of the short-listing methodologies ranging from $r = .32$ to .41, $p < .001$. The highest correlation was between the CPS and the SJT ($r = .41$, $p < .001$), as for both assessments, candidates chose from a list of pre-determined answers, whereas the AFQ requires candidates to produce

---

2 MRCGP is the Royal College of General Practitioners licensing examination in the UK.

3 SJT data was not available for 12% of candidates due to administration reasons.
written answers from open-ended questions. This level of overlap provides evidence of construct validity for each short-listing methodology and also suggests each has independent variance.

***INSERT TABLE 3 HERE***

In analysing associations between each short-listing methodology and SC scores, both uncorrected coefficients and coefficients corrected for restriction of range are reported. Uncorrected correlations underestimate the ‘true’ relationship between predictors and criteria when, as in this case, the predictors are used to select candidates to be included in the sample (to attend the next stage of selection) and thus restrict the variation in scores in the sample. The uncorrected correlation between the AFQ and the SC is $r = .26$, $p < .001$, indicating moderate predictive validity for a selection methodology. Similarly, the uncorrected correlation for CPS and SC is $r = .30$, $p < .001$ and the uncorrected correlation between the SJT and SC is $r = .46$, $p < .001$. When corrected for restriction of range each of these correlations increases to $r = .40$ (AFQ), $r = .44$ (CPS) and $r = .56$ (SJT) respectively. In comparing short-listing methodologies, the AFQ shows the weakest, but still moderately strong association, and the SJT shows the strongest association with SC performance.

Results show the strongest prediction of SC performance is a combination of all three measures (uncorrected $r = .51$, $p < .001$). Hierarchical regression analysis show all 3 measures have significant incremental validity over all other measures (see Table 3; $p < .001$). The SJT is the best single predictor (adjusted $R^2 = .213$, $p < .001$) and offers the most incremental validity over other methodologies. The predictive validity of the SJT is superior to both the CPS and the AFQ, but the CPS and AFQ are effective predictors.

Discussion

Main Findings

All three short-listing methodologies independently show sufficient predictive validity for selection purposes. Corrected correlations of above $r = .3$ between predictor and the criterion indicate a moderately strong level of prediction. Although there is scope for improvement, the existing selection process using the CPS test and AFQ is an effective methodology. The CPS content focuses on clinical knowledge and problem-solving and the AFQ focus on various non-clinical domains including empathy, integrity and coping with pressure. The results show that the AFQ adds significant incremental validity over the CPS test in predicting SC outcome. This might be expected as the AFQ targets non-clinical domains (empathy, integrity, coping with pressure) whereas the CPS focuses on clinical problem-solving. The SJT shows the strongest validity in predicting SC performance. This concurs with other studies using SJTs in College admissions procedures where this methodology is best suited to assessment of non-clinical domains.
Although each methodology independently predicted later SC performance, the most accurate overall prediction was using the newly designed SJT in combination with the other measures.

**Meaning of the study**

The degree of predictive validity is an important evaluative standard for any selection procedure. Although sufficient validity is achieved, the AFQs are relatively costly to implement in marking time (with 2 selectors scoring the responses independently). For large volume recruitment such as in GP in the UK, this is a relatively costly approach to short-listing. The SJT provides similar measurement properties as the AFQ but without these associated costs in administration and resources. When considering efficiency in addition to predictive validity, result show the optimum short-listing methodology is a combination of the CPS and the SJT. However, the SJT is significantly more expensive to develop and therefore there is only cost beneficial if there are large numbers of applicants.

**Conclusions**

The results have important implications for developing selection systems for large volume recruitment to optimise both efficiency and effectiveness. This is relevant to current proposals for selection into UK postgraduate training for all medical specialties in the use of machine-marked tests for short-listing purposes\(^{17}\). Future research should explore candidate reactions and perceptions of fairness \(^{5,18}\) and other crucial evaluative standards in judging the quality of selection methodologies. Further research could explore whether assessments are prone to coaching or practice effects \(^{14}\). Results at selection could be linked with other work-based assessment methodologies once trainees are in post to explore predictive validity in the long-term.
What is already known on this topic?

- In large volume recruitment, selection tends to use multiple stages to progressively decrease the applicant pool via application forms, short-listing tools and interviews.
- Little empirical validation is available on short-listing methodologies for selection into specialities at the post-graduate training level.
- Scoring application forms by trained assessors can be expensive in recruiter and administration time.

What this study adds?

- Clinical problem-solving tests, structured application form questions and situational judgement tests are all valid short-listing methodologies.
- In this setting the optimum battery of selection tools was a combination of a clinical problem-solving test and a new situational judgement test.
- A similar approach could be adopted for selection methodology for use in all medical specialities to enhance efficiency and effectiveness.
References

1. Modernising Medical Careers, Department of Health, UK. http://www.mmc.nhs.uk/
1. You are reviewing a routine drug chart for a patient with rheumatoid arthritis during an overnight shift. You notice that your consultant has inappropriately prescribed methotrexate 7.5mg daily instead of weekly.

Rank in order the following immediate actions in response to this situation (1 = Most appropriate; 5 = Least appropriate).

A. Ask the nurses if the consultant has made any other drug errors recently.
B. Correct the prescription to 7.5mg weekly.
C. Leave the prescription unchanged until the consultant ward round the following morning.
D. Phone the consultant at home to ask about changing the prescription.
E. Inform the patient of the error.

2. A patient has a history of drug abuse and has been in and out of prison over the past 3 years. He has been trying to obtain a prescription for diazepam. You are a Foundation year 2 doctor in general practice and he has been booked in to your surgery as an extra patient. He is a large man with many tattoos.

Choose the THREE most appropriate actions to take in this situation.

A. Give the patient a prescription for a small amount of diazepam.
B. See the patient and explain that you are not allowed to prescribe the diazepam as it is against the practice prescribing policy.
C. See the patient and explain that he needs to provide a urine specimen for a toxicology screen before you can prescribe diazepam.
D. Provide a prescription for the same dose and quantity that the last doctor provided.
E. Tell the reception staff that you will not see this patient as an extra.
F. Tell the reception staff that you will only see this patient as an extra if you have a chaperone.
G. Arrange for the patient to be reviewed in a day or two by a senior colleague.
H. Ask a senior colleague to see the patient today.
Table 1. Overview of three shortlisting methodologies

<table>
<thead>
<tr>
<th>Item Content &amp; Response Format</th>
<th>Clinical problem-solving test (CPS)</th>
<th>Structured application form questions (AFQ)</th>
<th>Situational Judgement Test (SJT)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
<td>The CPS evolved from an existing test of clinical knowledge developed by the North West region in the UK. Operational papers have 100 questions and a completion time of 90 minutes. A test contains a mix of extended match and single best answer formats covering the range of clinical areas defined by the UK training curriculum.</td>
<td>Seven AFQ items were drawn from an existing bank of items with each question targeting a competency domain including empathy, integrity and coping with pressure. An example question is, “Give an example of how you won the trust of a worried and sceptical patient. Describe what you did, and why, and the effect it had on both you and the patient? The questions cover aspects of both clinical and non-clinical domains. Candidates were allowed 2 hours to complete 7 questions and a 250 word limit per question. This assessment is completed under invigilated conditions.</td>
<td>The SJT is a newly designed instrument and consists of 50 questions of two different types (rating and multiple best answer response formats (see Figure 1 for examples). Twenty senior General Practitioners (4 female, 1 from a minority ethnic background), with experience both in training and in designing assessments, worked with 3 psychologists to generate a bank of questions. Questions target three domains including empathy, professional integrity and coping with pressure. In total, 186 items were written and divided into 4 pilot test forms of 50 items each (14 items were repeated in a second test form). Papers are machine-marked according to the agreed key.</td>
</tr>
<tr>
<td><strong>Scoring Process</strong></td>
<td>Papers are machine-marked according to the agreed key.</td>
<td>Each paper is scored independently by two trained assessors working from a previously validated scoring framework. Assessors were trained during a one-day workshop and calibrated to enhance reliability.</td>
<td>Papers are machine marked according to the agreed key.</td>
</tr>
</tbody>
</table>
Table 2. Summary of descriptives for all measures in the study

<table>
<thead>
<tr>
<th>N=463</th>
<th>Selection Centre</th>
<th>AFQ</th>
<th>CPS</th>
<th>SJT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.15</td>
<td>3.07</td>
<td>82.35</td>
<td>0.71</td>
</tr>
<tr>
<td>SD</td>
<td>0.46</td>
<td>0.26</td>
<td>7.5</td>
<td>0.57</td>
</tr>
<tr>
<td>Min,max</td>
<td>1.50, 4.00</td>
<td>2.1,3.7</td>
<td>64, 98</td>
<td>-1.24, 1.97</td>
</tr>
</tbody>
</table>

The SJT scores were standardised within trial group in order to combine scores across the different trial forms.
Table 3. Predictive validity results for the CPS, AFQ and SJT

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Adjusted R square</th>
<th>R square change</th>
<th>F (df1, df2)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPS</td>
<td>0.090</td>
<td>46.85</td>
<td>(1,461)</td>
</tr>
<tr>
<td>AFQ</td>
<td>0.066</td>
<td>33.43</td>
<td>(1,461)</td>
</tr>
<tr>
<td>SJT</td>
<td>0.213</td>
<td>125.74</td>
<td>(1,461)</td>
</tr>
<tr>
<td>CPS and AFQ</td>
<td>0.138</td>
<td>37.93</td>
<td>(2,460)</td>
</tr>
<tr>
<td>Incremental prediction of CPS over AFQ</td>
<td>0.074</td>
<td>39.62</td>
<td>(1,460)</td>
</tr>
<tr>
<td>Incremental prediction of AFQ over CPS</td>
<td>0.049</td>
<td>26.42</td>
<td>(1,460)</td>
</tr>
<tr>
<td>CPS and SJT</td>
<td>0.232</td>
<td>70.81</td>
<td>(2,460)</td>
</tr>
<tr>
<td>Incremental prediction of CPS over SJT</td>
<td>0.021</td>
<td>12.69</td>
<td>(1,460)</td>
</tr>
<tr>
<td>Incremental prediction of SJT over CPS</td>
<td>0.143</td>
<td>86.12</td>
<td>(1,460)</td>
</tr>
<tr>
<td>AFQ and SJT</td>
<td>0.235</td>
<td>71.87</td>
<td>(2,460)</td>
</tr>
<tr>
<td>Incremental prediction of SJT over AFQ</td>
<td>0.170</td>
<td>102.91</td>
<td>(1,460)</td>
</tr>
<tr>
<td>Incremental prediction of AFQ over SJT</td>
<td>0.024</td>
<td>14.35</td>
<td>(1,460)</td>
</tr>
<tr>
<td>CPS, AFQ and SJT</td>
<td>0.252</td>
<td>52.94</td>
<td>(3,459)</td>
</tr>
<tr>
<td>Incremental prediction of AFQ over CPS and SJT</td>
<td>0.022</td>
<td>13.38</td>
<td>(1,459)</td>
</tr>
</tbody>
</table>

Performance at the SC was used as the criterion variable, using the mean score across three simulation exercises in the SC. *All values were significant p<.001