Review of JAC Shortlisting Tools – Summary Report & Conclusions

Final Report

July 2018

Dr Máire Kerrin
Charlotte Flaxman

About Work Psychology Group

1.1. WPG are an independent evidence based consultancy that engages with organisations at a strategic level to translate the very latest research into innovative, effective and sustainable solutions. Our principal areas of business activity are services relating to selection and assessment, individual and organisational development, and evaluation. Our psychologists are uniquely skilled in these areas and are committed to working in partnership with clients to ensure our knowledge and expertise benefits the needs of the organisation.

1.2. Work Psychology Group has substantial expertise in the design, development and evaluation of selection and assessment tools and methodologies, including job analysis, role profiling and competency modelling. Of relevance to this review, previous work regularly involves the design and evaluation of psychometric tests and selection methods for high stakes selection, for example in medicine, finance and professional services. We have substantial experience and expertise in evaluating the impact of selection methods and working within the political and public interest similar to the context of this JAC work. A key part of this work includes monitoring of diversity through quantitative psychometric analysis of tests (e.g. group differences in performance) and quantitative and qualitative candidate feedback on the process (how they feel they were treated by the process).

Review of JAC Shortlisting Tools

Background & Context

2.1　The Judicial Appointments Commission (JAC) selection process is unique for multiple reasons. First there is a high volume of candidates for a small number of prominent positions, many of which are in effect 'lifelong' positions. Additionally, the roles are high profile with high stakes, and have a strong public interest. The selection processes used for these roles are subject to high levels of scrutiny, as there would be far reaching consequences for any poor quality appointment. While we regularly work with professions with similar high stakes and high levels of public interest (e.g. Medicine, Dentistry, Pharmacy), selection is often into a structured training programme. This is not the case for all JAC appointments, further emphasising the uniqueness of this selection process.

2.2　Designing a robust selection process for such a high stakes context is a complex and continual task. Important elements (e.g. role requirements, selection criteria, context for implementation and stakeholder acceptance) may change over time and it is important for evaluation and development to take place on an on-going basis. We welcome our involvement in this independent review, which sits within the context of the JAC continuing to aspire to ensure that their selection processes are as efficient, effective and fair as possible.

Project Overview & Objectives

2.3　In March 2018, we were commissioned by the JAC to review and evaluate the shortlisting tools for four large exercises. Specifically, the review contributes to the ongoing monitoring of any adverse impact[1] experienced by specific groups of candidates (Ethnicity/Age/Gender/Current Profession) and to evaluate the robustness of the shortlisting model and tools used.

2.4　The objectives of the evaluation were (1) to follow-up on reviews conducted in 2013 and 2015, (2) to review the effectiveness of the JAC shortlisting tools to make decisions that are fair and valid, (3) to evaluate the effectiveness of the stages of the selection process, and (4) to provide commentary targeting areas that require improvement.

2.5　A multi-method approach to the review and evaluation was developed consisting of a desk review, stakeholder consultations (in the form of telephone interviews), and quantitative analysis of shortlisting data.

Review Evaluation Criteria & Methodology

2.6　The review focused on four recent or ongoing large exercises (031 Deputy District Judge (Magistrates' Courts), 042 Recorder, 044 Deputy High Court Judge and 084 Fee Paid Judges of the First Tier Tribunal). The exercises used either a telephone assessment (TA) or qualifying test (QT) (or a combination of both) as shortlisting tools, with successful candidates progressing to the selection day. The results of exercise 084 are yet to be published and therefore the outcomes for this exercise are not for publication at this stage. The data collected as part of this review was evaluated against four criteria:

2.6.1　**Evidence of Fairness**: are selection criteria objective and valid, are the assessments standardised, how are outcomes monitored, what are candidate perceptions in terms of fairness.

---

[1] While we have reported on adverse impact in places, we cannot comment on the causality of this. Research in other professions on progression through selection, training and exams suggest a growing evidence and understanding of the multifactorial nature of causality for difference in performance between groups (https://www.gmc-uk.org/education/standards-guidance-and-curricula/projects/differential-attainment/what-causes-differential-attainment)

2.6.2 **Evidence of Validity**: does the selection tool measure what it claims to measure (e.g. face validity[2] and content validity[3]), what is the evidence for longer term measures of validity (e.g. construct[4] and predictive validity[5]).

2.6.3 **Evidence of Reliability**: does the selection tool measure consistently under varying conditions.

2.6.4 **Ongoing Development**: are arrangements in place to collect data for monitoring and review purposes, are appropriate actions in place to support validation and continuous improvement.

2.7 Desk review materials included design briefs, marking sheets, assessment materials, preparation materials, candidate feedback, and panel member briefing materials.

2.8 Telephone interviews were conducted with six judges and five internal JAC employees involved in the drafting process.

2.9 Psychometric analysis on shortlisting data was conducted on data available for exercises 044 and 084, looking at shortlisting outcomes broken down by demographic sub-groups (e.g. gender, age, ethnicity, current profession). The statistics for exercise 084 will be published in June 2019.

2.10 The overall review and evaluation focused on looking at benefits and limitations of the TA or QT as a shortlisting methodology after triangulating the findings from the various information sources, including selection data, telephone interviews and assessment materials.

Key Findings

Overview of Selection Process Strengths

2.11 Since the previous reviews of the JAC selection process in 2013 and 2015, there have been significant changes to the JAC shortlisting process and assessment, many of the changes embracing elements of best practice and standardisation of selection processes. In addition, the JAC have moved away from their selection methods specifically testing a candidate's knowledge of the law, and instead now focus on measuring a candidate's broader range of competence.

2.12 The outputs of the review indicate that overall, the JAC are following a best practice approach in the development of shortlisting materials (for both the TA and QT). This includes the input of judges' subject matter expertise to develop the content, stages of quality assurance and equality and diversity review throughout the process, input from external stakeholders (in the form of the advisory group) and the dry run process to pilot materials.

Findings Related to Evaluation Criteria

2.13 There are identified areas for improvement outlined in the conclusions and future consideration section. A summary of findings from the evaluation against the four criteria, is presented below.

2.13.1 **Evidence of Fairness**: Overall the review highlighted the use of objective and valid selection criteria based on competency frameworks. However, there was a strong focus on having knowledge of the 'correct thing to do' in relation to other more behavioural criteria. Whilst we recognise that there has been a shift away from testing legal knowledge, within the current format of some elements of the shortlisting process,

---

[2] Face validity refers to if a test or assessment method measures what it claims to measure. For example, a measure of face validity could be to ask candidates if they believe the assessment method completed is appropriate for the position they have applied for.

[3] Content validity refers to how well a test or assessment method measures the behaviour that it is intended to measure. For example, if a test should measure three different competencies, are all three being measured effectively as part of that test.

[4] Construct validity is the degree to which a test or assessment method measures what it claims to measure. Psychometric analysis (i.e. correlations) can be used to correlate a new method with a previously validated method which is anticipated to measure the constructs, to understand if the constructs are being measured accurately in the new method.

[5] Predictive validity is the extent to which performance on a test or assessment method can predict subsequent performance of work behaviour.

there may still be some benefit to those that have experienced certain types of situations in the past and therefore know what the correct thing to do is. The review also highlighted that the drafting process is extensive with stages to monitor standardisation, fairness and equality and diversity. Processes are in place to monitor candidate perceptions; however some additional questions could be included to further monitor perceptions of fairness.

2.13.2 **Evidence of Validity**: Involvement of judges throughout the drafting process provides good evidence of *face validity*, with materials being relevant to the job role, whilst also being mindful of ensuring the content is not so specific as to disadvantage certain candidates. Whilst there are a number of competencies identified to be measured in the materials, a strong focus on assessing competencies such as 'exercising judgement', 'possessing and building knowledge' and 'assimilating and clarifying information' could mean that other competencies previously identified as important are not being equally assessed. This results in restricted *content validity* (where the coverage of the competencies are narrow). Professional expertise regarding test item development was recognised by stakeholders in the consultations as limited and potentially having an impact on item content. Access to additional data would be required in the future to understand the *predictive validity* of the shortlisting methods (i.e. how does performance on the QT predict future performance either in the selection day, or later in the role).

2.13.3 **Evidence of Reliability**: The desk review provided evidence of an appropriate approach to standardising the development of the tools. However, there were variations in the structure of the scoring criteria used for the TA, such as, provision of detailed model answers compared to provision of a structured marking sheet. Further standardisation across the scoring criteria for the TA would mean greater confidence in competency scores and ensure candidates are being assessed in a consistent manner. Additionally, more specific assessor training (for both judges and panel members involved in assessing candidates) in best practice assessment principles and moderator training (e.g. for the scenario assessment Part 3 QT; focusing current TA training), will help with the overall reliability of the methods. Length of test, number and type of items within the QT will affect reliability and needs to be continually reviewed. It is recommended that for psychometric analysis on future exercises, data is analysed at the individual question level[6] to evaluate the reliability of that individual method. Whilst this is an important part of evaluating any test or assessment method, analysis will only provide details of the reliability of that individual test. Therefore this analysis should be conducted on methods following each exercise.

2.13.4 **Ongoing Development**: Processes are in place to collect data on the perceptions of the shortlisting methods and to monitor general perceptions of the process. Whilst it is clear from consultations that feedback is reviewed and changes are proposed for future drafting of materials, this is not explicit within documentation and therefore may not be transparent across all parties involved in the process. Additionally, whilst feedback is sought from candidates, additional questions could be developed to understand more about perceptions of fairness and relevance of shortlisting tools.

Psychometric Evaluation Summary

2.14 When reviewing the psychometric evaluation for exercise 044, no adverse impact[7] was found based on gender across all shortlisting stages. Whilst adverse impact was found for BAME candidates at the sift stage, no adverse impact was evident for BAME candidates in the telephone assessment or selection day. In relation to current profession, adverse impact was found across the sift and telephone assessment. When looking at disability, no

---

[6] This data was not available for the current analysis and report.
[7] Adverse impact refers to employment practices that appear neutral but have a discriminatory effect on a protected group. For the purposes of this report, adverse impact is calculated using the Adverse Impact Ratio (also referred to as the four fifths rule).

adverse impact was evident in the sift or telephone assessment. Sample sizes were too small to meaningfully investigate adverse impact based on current profession or disability within the selection day stage.

2.15   In addition, as part of our review, we have evaluated the impact of initiatives that the JAC and Judicial Office have implemented, on the success of candidates making it through various stages of exercises. Specifically, we evaluated the attendance at a JAC presentation or completion of a Work Shadowing Scheme for exercises 044 and 084. For exercise 044, the data showed that attendance at JAC Presentations was high among candidates, with over half of candidates attending a presentation. Whilst the overall number of recommended candidates for 044 was low, analysis indicates that more of the recommended candidates had attended a JAC presentation than had not. Whilst these initial evaluation findings were positive for both BAME and solicitor candidates that attended a presentation, no statistically significant differences in outcomes based on attendance were found.

2.16   In relation to Work Shadowing Schemes, approximately one third of candidates for exercise 044 participated in a scheme. Whilst proportionally, those candidates in minority groups that were recommended were in line with those that participated in the scheme, no statistically significant differences in outcomes was found based on participating in the scheme. As previously noted, we cannot report on exercise 084 as the results are yet to be published.

2.17   Whilst there were no statistically significant results found, these results are encouraging, in highlighting that many candidates are aware of the initiatives being implemented by the JAC, and are recognising the value of participating in these schemes. However, to understand the true impact of these schemes and understand whether participation can increase a candidate's likelihood of being recommended, further evidence and evaluation is necessary.

Conclusions & Future Considerations

2.18   The evidence generated through this review demonstrates a professional and concerted effort to develop shortlisting tools that are fair and valid. This is within the constraints of available resourcing and test development expertise that is internal to the JAC and the supporting Judiciary involved in the drafting process.

2.19   Whilst the review broadly concludes that the approach to the design of the tests in principle follow the paradigms of best practice, there are areas to develop that would support the fairness, reliability and validity of the tests, which we outline below in the future considerations.

2.20   Before looking at these considerations in detail, as part of the review we understand that the JAC is about to embark on an increasing number of large selection exercises per year. This will require an increasing number of tests to be developed, larger (potentially unknown) number of candidates that will need to be shortlisted, placing greater pressure on the shortlisting tools to reduce numbers for available interview slots. A major challenge that we observe with this based on the review evidence and our experience of high stakes, large scale selection test development is whether the current approach to test development will be able to keep pace and be sustainable (given the test design structure).

2.21   Given the increasing number of large exercises, the resource challenges in test development, and the evidence regarding multiple assessments to improve fairness, reliability, and validity of the tools, there are alternatives that could be considered to support the shortlisting process. Specifically, the use of technology within methods (e.g. non-live interviewing) could provide another approach for some exercises (e.g. entry level posts). This has a growing evidence base and use in high stakes professional contexts, with benefits for both the candidate and organisational (e.g. resourcing) perspective.

2.22   An additional challenge within this selection context is to identify and select into the role individuals that can do the job in the future (potential), rather than selecting only those who have already demonstrated they can do the job (experience). This poses particular challenges for the design of selection methods (e.g. tests,

interviews) to be able to accurately select and identify for potential. A possible way to explore this is through the Situational Judgement Test (SJT) element of the QT. Based upon recent discussions with the JAC, we understand that a particular point of interest is in reviewing the response format of the SJT element of the QT, as a way to look for potential within candidates. Developing SJT items that allow a judgement to be made regarding several different courses of action, rather than making a final decision about the one best thing to do in that situation, can allow the candidate to 'take a step back' from the situation and reflect on the appropriateness of certain actions or considerations before deciding how to respond to the scenario. By developing an SJT in this way, it removes the requirement for a candidate to make the decision on how to manage the situation (which may be testing knowledge of the correct thing to do) and instead provides insight into what a candidate may take into account when deciding how to respond. Through doing this, it will allow the JAC to assess the potential that a candidate has to do the job, rather than it assessing those that have done the job previously. Further, utilisation of different response formats will allow for greater differentiation amongst candidates to be made during the scoring process.

2.23   Given the high stakes, high volume (for some exercises) nature of the selection process, based on the findings that emerged from the review there are a number of areas for potential improvement that have arisen. The considerations have been presented in relation to short term considerations, medium term goals and communication/engagement. Whilst all of these considerations presented will help to make improvements to the overall shortlisting materials and drafting process, it is important to review these in line with internal timescales and deadlines for operational delivery to understand the feasibility of each of them.

SHORT TERM CONSIDERATIONS
(1) Design / Content
- C1: Streamline paperwork required for exercise development, including development of standardised templates and better integration of 'QA checklists' of design considerations
- C2: When developing content for the shortlisting materials, consider how to develop tests that enable broader assessment across required competencies
- C3: Consider how to ensure consistency and clarity across scoring and marking criteria
- C4: Consider re-designing how judges can best be utilised for their knowledge, experience and expertise in the drafting process, rather than as test content developers

(2) Use of Different Methods
- C5: Consider which exercises could benefit most from inclusion of a TA, and which (due to volume or level) are better suited to use of the QT

(3) Training
- C6: Consider the inclusion of enhanced guidance regarding best practice assessment principles to share with judges and panel members within assessor briefing sessions
- C7: Provision of training for those responsible for the development of content for the SJT component of the QT

(4) Ongoing Development
- C8: Review current approach for collecting candidate feedback, including overall process and the specific questions candidates are being asked.

MEDIUM TERM GOALS
- C9: Adaptation of the QT to assess a broader nature of competencies to help identify future potential in candidates. This may be particularly important for entry level posts
- C10: Alternative Shortlisting Option: Consider methods that can be introduced (or re-introduced) to the selection process to help with shortlisting large numbers. Whilst it is understood that the JAC have explored many different alternative shortlisting tools, it is important to note that

recently technology within this area has advanced considerably. Therefore, it may be useful to consider which methods to revisit as options to help with shortlisting large numbers and assessment of broader range of competencies e.g. non-live video interviews, or alternative approaches to SJTs (see C12)

- C11: Consider how the overall drafting process can be updated to streamline the process and reduce the resource and time needed, whilst retaining collaboration
- C12: Review alternative options for SJT response instructions format to allow for greater differentiation between candidates and a focus on overall judgement and candidate potential, rather than having the knowledge to make an appropriate decision

ENGAGEMENT / COMMUNICATION

- C13: If shortlisting processes do change, consider how to communicate these changes to all relevant stakeholders