Selection Interviews: Some Issues[*]

Antony Eagle, antony.eagle@adelaide.edu.au

0. Background

The interview is a central tool in our evaluation and selection of candidates for employment and (in some disciplines) admission. It should concern us greatly, therefore, that psychological research has consistently and repeatedly shown that interviewing is an unreliable indicator of candidates' future performance. It is also particularly vivid and influential compared to other, more reliable, sources of evidence. Involving interviews in the selection process is, unless carefully managed, nothing more than an expensive source of vivid noise. Since we don't intentionally wish to randomise our selection process, it is important that we are aware of the potential problems with interviews, so that we can take appropriate measures to avoid them.

I write as an end-user of interviews, rather than a subject matter expert. The literature is *much* vaster than what I survey here. Yet those parts of the literature I am aware of have convinced me that there are some pitfalls concerning interviews, and I hope at least that considering the evidence I offer below might prompt some reflection of interview practice that will hopefully improve the reliability and validity of the interview, for both our benefit and the candidates.

We are all diligent and responsible when undertaking interviews, and I certainly do not question anyone's professionalism or intend to cause offence. Certainly many readers will have much more experience in interviewing candidates than I have. I offer these comments in a constructive spirit: all of us, no matter how experienced, can benefit from reflection on our own practices, and it is a good idea to make as explicit as possible the principles we employ in making admissions decisions. For those of us who are less experienced, I hope this can help us to form good interview habits now.

In section 1, I review some of the psychological evidence. In section 2, I draw out some consequences for the way we do things that might prompt change in our procedures. In section 3, I give some positive suggestions that we can implement now to improve our current procedures.

1. The evidence

The most useful evidence for our purposes comes from evaluations of the *validity* of the interview process: that is, the correlation between the assessment of the candidate based on the interview and the true nature and potential of the candidate. In academic hiring, this might be a measure of the correlation between interview scores and some measure of performance in teaching or research. (Available measures of those attributes of course have many deep problems of their own.) We don't at present collect such data in any systematic way, particularly

[*] Draft. Contact me if you wish to cite. Comments welcome. This document represents my personal views and not those of the University of Adelaide.

when it comes to those we did not hire, but it would be useful evidence, at least to give us some empirical basis for our judgements of our own reliability.

A meta-analysis of studies on the validity of interviews (Marchese and Muchinsky 1993) makes a couple of noteworthy points that remain cogent:

- The overall correlation between interview score and actual performance (in employment, in their cases) was found to be 0.274 (variance 0.023); this is better than chance, but not much. They report that this 'is still below the average validity for cognitive ability tests and work samples as a method for personnel selection' (Muchinsky 1986). Even after significant correction of observed validity coefficients, the resulting validity of the interview was found to be 0.379 (variance 0.031). This is still only a medium positive correlation between interview score and actual performance.
- There is a weak negative correlation between length of interview and validity ($r = -0.29$, $p < 0.10$): the longer the interview, the less validity. This may be due to the fact that long interviews provide a large mass of potentially conflicting information; more pessimistically, the longer the interview continues, the more opportunity we give the candidate to rule themselves out without necessarily being indicative of their actual ability!

I think there is anecdotal support for some scepticism about interviews already. We're already aware that we react positively to some people, based on appearance, bearing, and other non-verbal cues. It would be unreasonable to expect that these 'good first impression' biases didn't occur in interviews. But we can do better than mere anecdote. For example, appearance appears to play a role (Posthuma *et al.*, 2002, 22–3), with overweight, poorly dressed, or unattractive applicants disadvantaged.

There is a very large literature on the existence of *implicit bias* in human judgment, by which I mean, biases that are reflected in someone's behaviour but which are not overtly verbally endorsed. Most people, even those who sincerely and consciously express a commitment to equality and fairness, are implicitly biased against historically disempowered groups, such as people from low SES backgrounds, black people, women, gay people, gender non-conforming people, and so on. (Obviously there is considerable intersection between these groups; and bias, implicit or explicit, surely plays a role in explaining why these characteristics of disadvantage cluster in the way that they do.) There is research that supports the natural expectation that implicit biases play a role in selection interviews, as elsewhere. Common stereotypes ascribed to males and females may affect interviewer's evaluation of candidates. One study (Schein 1973) asked 300 male managers to evaluate which of 92 adjectives best described (i) women in general (ii) men in general and (iii) successful middle managers. They found that the correlation between descriptions of managers and men ($r = 0.62$) was much higher than the correlation between descriptions of managers and women ($r = 0.06$). Other studies found that women interview better than men when applying for traditionally 'female' roles (Arvey *et al.* 1987). Together this suggests that expectations of traditionally 'male' and 'female' roles and abilities have a significant impact on interviewer's judgements of the suitability of candidates. We should be concerned at the sex distribution in admissions by different subjects.[†] Worryingly,

[†] Though of course we should also expect that this phenomenon will have shaped our possible candidate pool long before they come to us; prospective students are, however subtly, read as women and men in

Harris (1989) finds evidence of interviewer bias in favour of applicants of the same sex, and a disconfirmatory bias for applicants of the opposite sex. (Similar results have been found with applicant's racial and attitudinal similarities to interviewers: Posthuma *et al.* 2002, 5–6). If interviewers tend to favour candidates similar to themselves, obvious problems of access and the perpetuation of unrepresentative demographics in the university and disciplines will arise. When approaches to interview that minimise the impact of implicit biases are imposed, significant changes in hiring rates can occur. One famous study showed that when orchestral auditions were held behind a screen, so that panel members were ignorant of the gender of applicants, the chances for women musicians to proceed from a performance round to a subsequent stage increase dramatically when performances are anonymised (Goldin and Rouse 2000).

People's tendency to disregard sample size also plays a role in interview validity. Tversky and Kahneman (1974, p. 1125) showed that people overwhelming neglect differences in sample size when evaluating representativeness, and thus systematically fail to correct for the fact that small samples are far more likely to stray from the overall population. In the present context of interviews, this is evidence that we are prone to take a 30 minute interview to be equally representative of the candidate's abilities as is a much longer performance by the candidate (as reflected, for example, in their letters of recommendation or average performance in assessment over many years)—even though the interview is far more likely to give an inaccurate presentation of the candidates abilities (is more likely to be an outlier from the candidate's mean performance).

Interviewers are also subject to the 'anchoring and adjustment heuristic' (Tversky & Kahneman 1974), the tendency to make judgements by measuring a given experience against a pre-set anchoring expectation. If interviewers are given criteria with a high anchor, they tend to evaluate candidates more highly than had they been given a low anchor (Kataoka *et al.* 1997). We should expect that this heuristic is also operative when interviewers are asked to evaluate a candidate's performance at interview in light of their prior opinion, where a negative impression formed at pre-interview can provide a low anchor which in turn correlates with a low score at interview, thus reinforcing the prior opinion. Of course, the low interview score and low pre-interview impression may be correlated because the candidate is genuinely poor; but this level of correlation between interviews and prior measures is problematic if interviews are to provide another independent source of evidence.

People also seem to give undue weight to their own evaluations. Take a candidate who you rank highly on the basis of non-interview evidence. If that candidate performs badly at an interview for which you are present, that could easily rule that candidate out of the running in your mind. On the other hand, if you merely hear that the candidate has performed badly at another interview, you are less likely to revise your judgement based on the other evidence. (This explains something I'm sure you have often observed: the unwillingness of colleagues to devolve hiring to others, and desire to 'be involved' in the process. We do not explicitly regard our colleagues as worse than us: if we are rational, we shouldn't think their interview evaluations less reliable than our own.)

---

formation, and steered by family/friends/teachers in certain directions through gender-specific expectations encoded in culture.

All these studies and results cast doubt on the validity of interviews, though none is conclusive in showing there to be no positive benefit at all to interviews. I think the crucial problem comes from the *interview illusion*:

> that is, people's mistaken belief in their ability to predict, based on a brief conversation with someone, how they will evaluate this person in the future… The interview illusion persists even though we undoubtedly have many experiences of gradually changing our opinions of people as we get to know them better. We may think our reactions to a person after a brief interview predict how we will feel once we know them better because we simply do not realize that we no longer feel the way we had at the time of the interview; we believe we have always felt the way we do now. After many such experiences of mistakenly recalling our initial reactions to a person as similar to our current ones, we may come to believe that the reactions we experience when we first meet a person are highly predictive of how we will feel toward this person once we get to know him or her well. (Kunda 1999, 179–80)

The interview illusion is an instance of a very general tendency we have to exaggerate the consistency between the past and the present (the consistency illusion).[‡] Moreover, our mistaken faith in the predictive power of interviews persists even though we have experience of their unreliability. In one study (Kunda and Nisbett 1986), people were told that a psychiatrist had interviewed applicants to the Peace Corps, and we asked to estimate how well these interviews could predict the applicant's performance. People judged these predictions as quite accurate: they estimated the correlation between the interview and the candidate's performance at about .60. In fact, the actual correlation was about .07; we can substantially overestimate the extent to which performance can be predicted on the basis of interviews.

I can imagine this kind of response: 'I am not subject to these effects, because my track record shows my use of interviews has managed to select better from worse candidates'. It is, unfortunately, likely that this kind of response is an instance of the *choice-supportive bias*: a tendency to attribute, both correctly and incorrectly, more positive features to an option we selected than to a rival we did not (Mather *et al.* 2000).[§] So I suspect even our positive evaluations of our current practice is unlikely to be much good, without considerably more control over the evidence we use in making these posterior evaluations of our interviewing ability. Particularly in academic hiring, particularly these days, almost all candidates would turn out to be reliable, interesting, and productive colleagues. We shouldn't be overly impressed by the fact that we've picked good colleagues when the pool is so strong to begin with.

In sum, interviews provide unreliable data, and we are not capable of appropriately weighting that unreliable data because of the interview illusion and the choice-supportive bias. While

---

[‡] And is in fact probably a consequence of the fact that most of our memories are reconstructed from a very limited set of facts recorded at the time, and fleshed out by adding our current knowledge of regularities and facts we now believe to have obtained then.

[§] The hypothesis is that this bias might help us feel better, by reducing regret for options not taken; and interestingly it seems to get stronger as we age.

these results are presumably not exempt from the reproducibility crisis that has enveloped social psychology, perhaps particularly the results on implicit bias, still there is sufficient evidence pointing in the same direction to justify some confidence in these results.

2. Consequences

We should not expect that, psychologically speaking, academics are any more immune to these biases and failings of the interview process than anyone else. As such we should be very suspicious of the evidential value of interviews, but the interview illusion means that we very likely will not be. The safest course, then, should be to avoid interviews altogether: they are likely to provide noisy unreliable data and we will not be able to avoid relying on it. This is the course I prefer, and if I were designing a candidate selection system for academic hiring or admissions from scratch, I would give interviews next to no role.

Of course, that is not our position, and we have to consider what can be done to improve the validity of interviews in our present system. It is true that many departments are moving to systems in which non-interview data plays an important role in hiring and admissions decisions, but even there the interview is often the decisive factor when choosing between two close candidates. There are interestingly different issues in the cases of academic hiring and admissions interviews.

Academic Employment In academic hiring of tutors and lecturers, we can thankfully avoid interviews fairly easily. There is a large amount of data available about prospective hires that does seem to correlate well with performance and ability.

- *Responses to selection criteria.* This gives candidates a chance to marshal evidence relevant to their performance of the actual tasks of the job. It can serve the same function as an interview without the distorting effects of the interview situation – there are no follow-ups, no time pressure, no 'performance' aspects. The evidence on which candidates typically rely may be problematic—particularly the use of student evaluations in teaching portfolios[**]—and there is no guarantee that candidates will make use of comparable sources of evidence. (Obviously everyone will omit anything making them look bad.) But so long as we expect all candidates to be presenting images of themselves which are biased in the same direction, it may still be possible to extract useful comparative data, particularly if guidance is given to candidates about what kind of data the selection panel desires to have access to—for example, we could require a teaching portfolio containing detailed syllabi including lecture notes and assignments for a standard course the candidate would be expected to teach.
- *Written work.* This is a particularly reliable indicator of future research performance simply because producing written work is what that future performance will consist in. Consideration could be given to anonymous review of written work, given demonstrable biases in evaluation of candidates based on stereotypes applied in virtue of candidates names and other identifying information.

[**] Standard omnibus evaluations of teaching effectiveness are highly unreliable (Stark and Freishtat 2014). Even in the best case of reliable student evaluations, unfairness commonly results when student evaluations are the principal source used to evaluate instructors (Esarey and Valdes 2020).

- *Research and teaching presentations* will be less reliable – they will exhibit sample bias and implicit bias problems at the very least. Nevertheless, because they involve evaluating the candidate doing a task which will actually be part of their job description, they are more likely to be a good indicator than performance on tasks, like being interviewed, that candidates won't be required to perform as part of their job.
- *Academic references.* These too can be reliable, both because they are generated from a large sample of experiences with the applicant, and because many referees have a strategic interest in being honest: chances are we will come across future letters by that same person and dishonesty in the past will thus taint future experiences (cooperative behaviour emerges when entities have repeated encounters). Of course the bias of letter writers will come through in the letters too, so the source is somewhat tainted: and there is some evidence of systematic biases in letters of recommendation (Moss-Racusin, *et al.*, 2012).

Doing away with interviews for academic posts could be easily done; other institutions have managed to do so fairly successfully (Princeton Philosophy for example). But many departments will be hesitant to abandon in-person interviews, because they will be hesitant to abandon teaching and research presentations. It is likely that reliability of hiring will increase if departments move away from selection interviews, even if other components of the traditional interview/campus visit remain in place.

3. Improving Interview Validity and Reliability

But for many departments, internal politics and external HR requirements will make moving away from selection interviews impossible. But interviews can still be improved. There is fairly good consensus on what we can do to improve their reliability and validity. The main recommendation is that we should move to *structured* interviews (Campion *et al.* 1997). This means ensuring that each interview has the same structure, and that interviewers should have fairly narrow parameters within which to work. In practice, here are some ways to implement this:

- The main one is that **different candidates should be asked the same questions**, and given the same opportunities to display the core knowledge and skills we are looking for. This might seem obvious, but it is all too easy to end up altering our questions through the day (we get bored of the same phrasing, the candidates don't seem to be 'getting' our questions, etc.). All the evidence says that valid comparisons of candidates require a genuinely similar stimulus for their responses. This point is basic to any system of evaluation, and so extends also to examination and assessment: when different students sitting the 'same' test – or facing the same interviewers – can produce non-overlapping sets of answers, it's not clear that they are in any meaningful sense sitting the same test, and this makes it difficult to rank students with respect to their performance on that test: '… unless students are answering the same questions that measure the same outcomes, […] the inferences made about student ability are less valid' (Piontek 2008, 2).
- Give **candidates questions in advance**. We are looking for the strongest answers to our questions, not who is able to give the most satisfactory answer off the cuff. Unless for some reason the ability to give responses extemporaneously is a selection criteria, it

should probably not be made an implicit requirement through its involvement in interview responses.

- **Give interviewers a common set of explicit criteria** to ensure that (i) all interviewers are measuring the same qualities in applicants; and (ii) that the interview is actually measuring the qualities that are being looked for in the candidate (Posthuma *et al.*, 2002, 38–40)
- **Limit follow-up questions and limiting prompting of the candidates.** We can easily end up favouring one candidate over another equally knowledgeable candidate if our follow-up questions lead the first candidate's interview on to topics where the candidate has good knowledge, but the second candidate's interview (for purely contingent reasons) doesn't go in that direction and they do not get to display the same knowledge. It can be awkward to watch a candidate squirm, but prompting them will only distort the evidence (even if we suspect they do know the answer, we need to question the basis for our assumption, and question why we haven't made that assumption about other candidates who failed to correctly answer the question).
- **Restrict questions by the candidates**: don't let them distort the interview by attempting to bring it around to their favoured topics (and don't let them mess up an interview with a misguided attempt to impress).
- **Rate each answer independently**, rather than trying to give an overall mark that gets the 'gist' of the interview. The *recency effect* – the tendency for more recent events to be better remembered than those more distant in time, even over very short time periods, when we are tasked to recall them in free order (Atkinson and Shiffrin 1971) – probably means that we don't correctly recall the responses to earlier questions as well as responses to later ones, making a single overall evaluation less reliable than individual evaluations of each answer.
- **Restrict discussion between interviewers before assigning interview marks**; this is more likely to preserve the full range of opinion rather than grey out all the intra-interview variation in assessment. It's also sensible in light of the anchoring and adjustment heuristic, since prior conferral will likely provide an anchor that could easily be set in each case by a more dominant member of the interviewing team.
- **Detailed notes on candidate performance should be kept**, to facilitate later discussion on candidates at the end of the entire interview process (otherwise primacy and recency effects are likely to distort judgements).

All of these recommendations are made to improve validity, particularly to reduce the contamination by irrelevant information. While implementing these suggestions would require changes in the way we proceed in interviews, I hope the importance of improving our standards in this area to be as good as they possibly can is clear.

References

Arvey, R. D., H. E. Miller, R. Gould and P. Burch (1987), "Interview validity for selecting sales clerks". *Personnel Psychology*, vol. **40**, pp. 1–12.

Atkinson, Richard C. and Richard M. Shiffrin (1971), "The Control of Short Term Memory". *Scientific American*, vol. **225**, pp. 82–90.

Campion, Michael A., David K. Palmer and James E. Campion (1997) "A Review of Structure in the Selection Interview". *Personnel Psychology*, vol. **50**, pp. 655–700.

Esarey, Justin and Natalie Valdes (2020) "Unbiased, reliable, and valid student evaluations can still be unfair". *Assessment & Evaluation in Higher Education*, vol. **45**, pp. 1106–1120, doi:10.1080/02602938.2020.1724875.

Goldin, Claudia, and Cecilia Rouse (2000) "Orchestrating Impartiality: the Impact of 'Blind' Auditions on Female Musicians". *American Economic Review*, vol. **90**, pp. 715–41.

Harris, M. M., (1989) "Reconsidering the employment interview: A review of recent literature and suggestions for future research". *Personnel Psychology*, vol. **42**, pp. 691–726.

Kataoka, H.C, Latham, G. P., and Whyte G., (1997) "The relative resistance of the situational, patterned behavior, and conventional structured interviews to anchoring effects". *Human Performance*, vol. **10**, pp. 47–63.

Kunda, Ziva (1999) *Social Cognition: Making Sense of People*. MIT Press: Cambridge MA.

Kunda, Ziva and Richard E. Nisbett (1986), "The Psychometrics of Everyday Life". *Cognitive Psychology*, vol. **18**, pp. 195–224.

Marchese, Marc C., and Paul M. Murchinsky (1993), "The Validity of the Employment Interview: A Meta-Analysis". *International Journal of Selection and Assessment*, vol. **1**, pp. 18–26.

Mather, Mara, Eldar Shafir, and Marcia K. Johnson (2000), "Misremembrance of Options Past: Source Monitoring and Choice". *Psychological Science*, vol. **11**, pp. 132–8.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman (2012), "Science faculty's subtle gender biases favor male students". *Proceedings of the National Academy of Sciences*, vol. **109**, pp. 16474–9.

Muchinsky, Paul M. (1986), "Personnel Selection Methods". In C. L. Cooper and I. Robertson (eds.) *Review of Industrial/Organizational Psychology*, Wiley: London.

Piontek, Mary E. (2008), "Best Practices for Designing and Grading Exams", *CRLT Occasional Papers* **24** , University of Michigan Center for Research on Learning and Teaching, http://www.crlt.umich.edu/publinks/CRLT_no24.pdf.

Schein, V.E. (1973) "The relationship between sex role stereotypes and requisite management characteristics". *Journal of Applied Psychology*, vol. **57**, pp. 95–100.

Stark, Philip B. and Richard Freishtat (2014) "An evaluation of course evaluations". *ScienceOpen Research* doi:10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Tversky, Amos and Daniel Kahneman (1974) "Judgement under uncertainty: heuristics and biases". *Science*, vol. **185**, pp. 1124–31.

Version of 2021-04-27.