
Dewberry, Chris and Jordan, Deborah (2006). Do consensus meetings undermine the validity of assessment centres? *In: Division of Occupational Psychology Annual Conference, 11 – 13 January 2006, Glasgow. Proceedings.* Leicester: British Psychological Society

---

Do Consensus Meetings Undermine the Validity of Assessment Centres?

Chris Dewberry and Deborah Jordan
Birkbeck, University of London, and Ernst and Young

Abstract

In this study the effects of latent-informal processes operating in assessment centre consensus meetings is investigated with a combination of qualitative and quantitative methods. Non-participative observation is carried out in several consensus meetings, and auditory recordings made in three of these. In an analysis of the transcript of a consensus meeting in one organization, evidence is found for several latent-informal processes. These include active attempts by assessors to persuade other assessors, and the group facilitator, to appoint candidates; the use of assessors' general impressions of candidates in this persuasion process; and the active use of power derived from an assessors' relative seniority in the organization. Evidence consistent with the use of seniority-derived power is also found in a quantitative analysis of the selection decisions made in consensus meetings about 413 candidates. The results of the study are considered in relation to the practical utility of consensus meetings, and it is concluded that the use of such meetings is difficult to justify.

Meta-analytic studies indicate that the predictive validity of assessment centres (ACs) involving consensus meetings is less than might be expected, and on a par with unstructured interviews (Gaugler *et al.*, 1987; Schmidt & Hunter, 1998). One possible reason for this is that valuable information about candidates, derived from AC exercises, is degraded or lost during the consensus meetings typically used to make final decisions about candidates. Whilst there has been some speculation about the role of social psychological processes in consensus meetings, and the effect that these may have on decision-making (Herriot, 2003), this is the first study to examine evidence for the existence of these processes, and to examine their nature.

The Use of Assessment Centres

Assessment centres are widely used for personnel selection, promotion, and development planning. Gaugler et al. (1987) report that several thousand organizations utilize them in the United States, and a survey by Keenan (1995) found that 44% of British organizations engaged in graduate recruitment use them. Although the format of these centres varies so much that it may be argued that the "typical" centre does not exist (Bender, 1973), commonly the performance of a group of 10 to 20 assessees is evaluated in relation to a finite number of pre-specified dimensions. This evaluation takes place over one or two days, is carried out by multiple trained assessors, and involves several assessments or "exercises". A survey of assessment centre practices in the United States (Spychalski *et al.*, 1997) found that the most common assessments were in-basket exercises (81%), interviews (57%), analysis problems (49%) leaderless group discussions (43%), fact-finding exercises (38%) and skills and ability tests (31%). When the exercises have been completed there is usually a final consensus meeting where information about the candidate is brought together, and a decision is made about whether or not offer him or her a post.

The Justification for Assessment Centres

In comparison with other common selection techniques, assessment centres are relatively complex and consume significant organizational resources (Muchinsky, 1986). They are also expensive, the cost per employee during the 1980's estimated at between $50 and over $2000 (Cascio, 1986). Nevertheless, the use of ACs often appears justified. The use of job-relevant exercises (e.g. in-basket exercises, and analysis problems) promote high face-validity, and research carried out in the 1970's and early 1980's indicates good levels of criterion-related validity (Byham, 1970; Cohen *et al.*, 1977; Howard, 1974; Muchinsky, 1986; Thornton & Byham, 1982). This view is captured by Cascio (1986) who writes that

> The features of the assessment centre method - flexibility of form and content, the use of multiple assessor techniques, standardization methods for interpreting behaviour, and pooled-assessment judgements - account for the successful track record of this approach over the last few decades…Both minorities and non-minorities, men and women, acknowledge that the method provides them a fair opportunity to demonstrate what they are capable of doing in a management job. (p.266)

However, more recent meta-analysis of the many criterion-related validity studies carried out on ACs reveals a less positive picture. First, in absolute terms, the overall ratings given to candidates in assessments centres only explains about 14% of

their subsequent job performance (Gaugler et al., 1987). Second, and more tellingly, the level of criterion-related validity shown in the overall ratings given in ACs is no higher than for the exercise types of which they are composed. The situational exercises used in ACs, such as leaderless group discussions, and in-basket exercises, are forms of work sample, and these have a validity coefficient of .54 (Hunter & Hunter, 1984). Meta-analyses indicate that of other common exercises in ACs, structured interviews have a validity coefficient of .51 (McDaniel *et al.*, 1994), as do tests of general mental ability (Hunter, 1980). All of these figures compare very favourably with the validity coefficient of .37 obtained by Gaugler et al. (1987) for ACs, a figure on a par with the validity coefficient of .38 found for unstructured interviews (McDaniel et al., 1994). This is paradoxical: sophisticated and expensive ACs often incorporate structured interviews, cognitive ability tests, and work samples, yet when information about candidates on these measures are combined in ACs, the evidence suggests that they are less predictive than when used independently.

The Design and Implementation of Assessment Centres

There are several explanations for the lower than expected predictive validity of assessment centres. One is that whilst ACs are sound in principle, in practice they are often poorly designed and executed. This is discussed by Caldwell, Thornton, & Gruys (2003) who identify 10 errors associated with the design and use of ACs: poor planning, inadequate job analysis, weakly defined dimensions, poor exercises, the absence of pre-test evaluations, the use of unqualified assessors, inadequate assessor training, inadequate candidate preparation, sloppy behavior documentation and scoring, and the misuse of results. The implication here is that if these deficiencies are corrected, the criterion-related validity of ACs will necessarily improve.

Although these deficiencies can be expected to undermine the validity of ACs, Spychalski et al's (1997) survey of assessment centre practices in the United States indicates that in practice the centres are generally well designed and carried out. For example, consistent with the *Guidelines and Ethical Considerations for Assessment Centre Operations* (Task Force on Assessment Centre Guidelines, 1989) they found that 93% of ACs were based on job analysis. Assessor training was generally comprehensive and lasted about four days on average, and assessors used multiple methods to record their observations of assesses (e.g. 95% took notes, 41% used check lists, and a further 24% used behavioural observation scales). Most ACs were evaluated for reliability and validity, and in 80% of cases a periodic review, designed to ensure that the assessment centre continued to be appropriate for the target job, was carried out. Whilst there are undoubtedly deficiencies of one sort or another in the ACs run by many organizations, the picture suggested by this survey is that organizations generally operate acceptable levels of good practice in both design and execution.

Construct Validity and the Exercise Effect

A second explanation for the lower than expected validity of ACs may be found in the considerable research evidence on their construct validity. The correlation between different dimensions rated within the same exercise is typically greater than the correlation between ratings of the same dimension across several exercises (Archambeau, 1979; Neidig & Neidig, 1984; Sackett & Dreher, 1982; Shore *et al.*, 1990). This well-established phenomenon is known as the *exercise effect*. The implication of the exercise effect is that the criterion-related validity of ACs is not based, as assessment centre architecture assumes, on the ability of assessors to

discriminate between the performance of assessees on different dimensions within the same exercise, and for this reason ACs are seen as lacking construct validity. If the predictiveness of ACs relies on assessors to discriminate across several dimensions within an exercise, the failure to be able to do so may help to explain why the criterion-related validity of AC's is lower than might be expected.

However, recent research suggests that the criterion-related validity of AC's does not depend on the ability of assessors to discriminate in this way. Researchers (Highhouse & Harris, 1993; Neidig & Neidig, 1984; Schneider & Schmitt, 1992) have proposed that the exercise effect is not a result of rating deficiencies, but instead reflects real performance differences in the assessees across exercises. That is, assessees tend to do better in some exercises than others, and when they do well (or badly) they tend do well (or badly) in relation to all dimensions.

Lance et al. (2000) go further, presenting findings which suggest that overall assessment centre ratings are a function of two factors: (a) general assessment centre performance (i.e. the general performance of assessees across all exercises), and (b) a situation-specific performance factor (i.e. the performance of assessees in specific exercises). Supporting this proposition, they found that both the general AC performance factor, and the specific exercise factors, each explain a unique proportion of the variance in overall assessment centre ratings. These findings are confirmed by Lance, Foster, Gentry, and Thorenson (2004), who conclude that the assessment of candidates in AC's proceeds in two stages: An assessor forms an overall impression of a candidate's performance, and this general impression then drives the ratings which he or she assigns to the candidate on the prescribed, formally defined, dimensions.

If the criterion-related validity of ACs depends not on the ability of assessors to discriminate between the performance of assesses on different dimensions within an exercise, but rather on their ability to evaluate the assessees' overall performance both within each exercise, and across all exercises, the exercise effect provides a less potent explanation for the disappointing validity of ACs.

Consensus Meetings

A third explanation for the lower than expected predictiveness of ACs, explored in this article, is that valuable information about candidates is corrupted and lost, that erroneous, irrelevant, and misleading information about them is introduced, or that other dysfunctional processes and events occur, after the assessors have rated them, and that this occurs in the consensus meeting (sometimes called the "wash-up" or "wrap-up" session) held at the end the assessment centre. Surveys suggest that these consensus meetings are used in 84% of ACs in the United States (Spychalski et al., 1997) and 96% of those in the UK (Boyle *et al.*, 1995), and it has been suggested that they and consume 25% to 33% of the entire assessment centre cycle (Gilbert, 1981). Despite the popularity of these meetings, very little research has been carried out on them (Lievens & Klimoski, 2001).

What happens in consensus meetings? Although their format, like other aspects of ACs, varies across organizations, assessors usually meet after the exercises have been completed and consider each candidate in turn. In this chaired session (the chair is sometimes referred to as a facilitator), all the scores given to a candidate are presented, possibly in a matrix with exercises in rows and dimensions in columns, or vice versa. Assessors may provide a justification for each of their scores by presenting evidence based on their observations of the candidate, and may also

attempt to reconcile any differences in the ratings given to a candidate on a given dimension through discussion.

If relevant information about candidates is degraded or lost during consensus meetings, this would provide an explanation for the paradox that ACs have lower criterion-related validity than the exercises of which they are composed. The possibility that processes taking place within the wash-up are responsible for degrading the performance of ACs is highlighted by research indicating that "actuarial decisions", based on the weighted arithmetic sum of the ratings of assessees across dimensions, are equally or more valid than "clinical" ones involving a discussion of the assessees (Feltham, 1988; Herriot, 2003; Jones *et al.*, 1991; Pynes & Bernardin, 1989). The results of recent meta-analytic validation study by Authur et al. (2003) are also consistent with this argument. Authur and his colleagues collapsed the dimensions used in 34 criterion-related validity studies to six overall dimensions: consideration/awareness of others, communication, drive, influencing others, drive, organizing and planning, and problem solving, and then regressed job performance on these. They found that a model based on four of the six dimensions explained considerably more variance in job performance (20%) than overall assessment centre ratings (14%). If ratings on assessment centre dimensions are more predictive than overall assessment centre scores, the implication is that a simple weighted model of these scores is more effective at predicting job performance than the outcome of the discussion which takes place during consensus meetings. Put simply, rather than enhancing the validity of ACs, the implication of Authur et al.'s study is that the consensus session is undermining it.

What is going on in consensus meetings that might have a negative impact on the criterion-related validity of ACs? Herriot (2003) discusses this issue in relation to social psychological theory and research on group processes, and suggests that social identity and self-categorization theories (Tajfel & Turner, 1986; Turner, 1986) may be relevant. He argues that where organizational identity is salient, assessors may be more likely to select candidates who fit the organizational prototype of the ideal employee. Herriot also points out that we might expect social conformity processes to take place in consensus meetings. Conformity is known to result from both normative and informational influence. In the case of normative influence, the individual conforms because he or she needs to feel accepted by the group, or fears rejection (Deutsch & Gerrard, 1955; Wood *et al.*, 1994). With informational influence, individuals' judgements are based on information obtained not only directly through their own experience, but also indirectly through the opinions of other group members (Bishop & Myers, 1974; Kaplan & Miller, 1987; O'Reilly & Caldwell, 1979; Turner *et al.*, 1989).

Where there are differences in the experience, age, and/or status of people in the consensus meeting, we might expect these to intensify conformity effects, as assessors with greater power are likely to influence the ratings of others (French & Raven, 1959; Raven, 1965, 1993). As well as having power to reward or punish juniors who disagree with them (e.g. via the organizational promotion process) juniors may view their seniors' greater experience of personnel assessment as meaning that their judgments should carry more weight than their own. Thus more junior staff may defer to senior ones as a consequence of the combined effects of normative and informational influence.

The idea that such social influence processes are taking place in consensus meetings suggests that these meetings are characterized by both formal/explicit processes, and informal/latent ones. As there are no published or generally agreed

procedures for consensus meetings (Lowry, 1997), the explicit processes probably differ across consensus meetings. However, we might expect the training generally provided to those attending ACs (Spychalski et al., 1997) to include some explicit guidance on how to behave in consensus meetings, probably stressing the importance of rational processes, such as requiring assessors to provide evidence for their ratings, using information from various sources to test competing hypotheses about candidates, and combining information about a candidate on a particular dimension by adding together their obtained scores across several relevant exercises.

Examples of informal and latent processes running alongside these explicit processes would be the effects of conformity, including normative and informational influence, and the effects of power differentials between assessors. Although the small amount of empirical research on consensus meetings is patchy in focus, lacking in an underlying theoretical framework, and generally over 20 years old, that which is available supports the likely existence of such latent processes. Sackett and Wilson (1982) studied consensus meetings in a single organization, and found that when disagreement amongst raters required a consensus discussion, assessors sometimes varied in the degree of influence that they exerted. The existence of differentials in the influence of assessors is also supported by a laboratory simulation study of consensus meetings, carried out by Klimoski et al. (1980), in which they found that chairpersons exerted a relatively large influence on ratings, and by the results of a field experiment by Lowry (1992) which indicated that when there are differences in assessor seniority, and it is possible for assessors both to announce their scores publicly and to present arguments against the scores of other assessors, this can have a disproportionate impact on the overall ratings given to candidates.

In addition to these latent processes in interpersonal and group influence, there is also evidence for informal/latent processes in the way that assessors process information about candidates. A study by Russell (1985) indicated that when observing candidates in exercises, assessors initial ratings were dominated by a single general factor, and that this was usually based either on a candidate's perceived interpersonal skills, or on their perceived problem-solving skills. Similarly, and as mentioned earlier, Lance et al. (2004) found evidence that assessors begin by forming a general impression of candidates, and that this overall impression then drives their ratings of the candidates on specific, formally defined, dimensions.

To summarise, the implication of the theory and research discussed above is that the formal and explicit procedure which assessors have been trained to employ in consensus meetings is shadowed by a set of informal and latent processes, in which assessors, having derived a positive or a negative general impression of a candidate, actively seek to persuade others whether or not this candidate should be appointed, and where possible draw upon their status and position power within the group to do so.

The primary contribution of the present research is to examine, for the first time, whether evidence of all aspects of these informal influence processes – the formation of general impressions of candidates, the use of active persuasion as to whether a candidate should be appointed, and the influence of power relations, can be identified in a single, modern, and well-run assessment centre utilizing well-trained assessors. The principal focus of the research is to establish whether these processes occur in practice, and to examine the way in which this takes place, rather than to quantify the degree or frequency with which they occur. To this end, both qualitative and quantitative methods are employed. The qualitative technique involves auditory recordings of real AC consensus meetings, and the quantitative technique a statistical

analysis of the associations between scores given to candidates in AC exercises and the final selection decisions made about them. An advantage of the qualitative technique is the potential to examine, in detail, the nature of informal/latent processes such as active persuasion. However, one of the processes of interest here, the influence of the seniority and power of assessors on the selection process, is difficult to demonstrate unambiguously with transcripts of auditory recordings. Consequently, in this case data from the transcripts is combined with the results of the quantitative analysis. This affords a more rigorous examination of the evidence for the influence of power in consensus meetings than would be possible by examining the qualitative data in isolation.

The following propositions are examined: (a) that in consensus meetings assessors make active attempts to persuade others that a candidate should, or should not, be selected; (b) that during this persuasion process assessors draw not only upon information relating to the performance of candidates on formally agreed dimensions, but also on their general impressions of candidates; and (c) that assessors with relatively high levels of power, derived from their comparative seniority in a consensus meeting group, utilize this to persuade other assessors whether or not to select candidates.

Method

Qualitative Analysis

In order to examine the possible co-occurrence of explicit and latent processes in consensus meetings, I carried out non-participative observation of four assessment centre consensus discussions in three organizations: a multi-national financial organization, a medium-sized software company, and a division of the British army. Auditory recordings were made in three of these meetings: two in the financial organization, and one in the software company. These recordings were later transcribed. The researcher then examined the transcripts, until a clear example of a particular process could be identified. In all cases, appropriate examples were found in the first consensus meeting which took place in the financial organization, and all of the following discussion extracts are taken from this. The meeting concerned took place in an assessment centre designed to select graduates. The organization had approximately 12,000 applicants each year for about 400 graduate posts. The selection process began with an application form competed either on paper or on the Internet. When evaluating the application those involved in this initial stage of selection focused primarily on each applicant's academic record at school and university. About 3000 applicants were then invited to the second stage of the selection process in which they attended a first interview with a relatively junior member of the organization, and completed a cognitive ability test.

Approximately one third of these applicants were then asked to attend an assessment centre in which they were evaluated against eight dimensions: leadership potential, personal drive, interest in business and commerciality, practical intellect, commitment to clients and results, responsiveness to change, teamwork, and communication skills. The exercises used to evaluate the applicants were a case study (for which they were given 90 minutes), a leaderless group discussion (40 minutes) and a second interview (approximately 60 minutes). The second interview was always carried out with a senior member of the organization who, if the candidate was accepted, would lead the team in which they worked. The assessors for the other two exercises were comparatively junior members of the organization. All assessors had a minimum of two days training. Assessors who evaluated the candidate gave a rating

of 1 (poor), 2 (marginal), 3 (good), or 4 (strong) against each of the dimensions considered relevant to the exercise they observed or conducted.

At a subsequent consensus meeting, facilitated either by a senior member of the human resources department or by a member of the graduate recruitment team, each candidate was discussed in turn. Assessors announced the marks they had given to the candidate for each dimension, and these were written up on a board, observable to all, by the facilitator. Immediately after announcing their marks, assessors were required to present evidence for each one. During this process assessors had access to the candidates performance at the second stage of the selection process: their scores on the cognitive ability test (verbal, numerical, and spatial ability) and the scores they had obtained in the first initial interview.

Evidence of the Formation and Use of General Impressions of the Candidate

The following extract is from a second interviewer announcing the marks given to a particular candidate and providing supporting evidence for them.

Extract 1

I had 4, 4 for responsiveness to change, er 3 for communication skills and 4 for leadership. And I thought she was excellent ...one of the best candidates I've interviewed in a long time. I see what you mean about talkative and struggling to get to the point though which is why I marked her down slightly on communication skills. She is Japanese, she came to the UK when she was 15 or 16, and she has the most fluent English of any Japanese person I have ever interviewed. Her English is just impeccable, couldn't fault it. She's at Oxford, she I mean her biggest change was coming to the UK and assimilating, she does have Japanese parents, and I thought one of her parents might be English which would account for the English but both are Japanese. She is, ah, tremendously driven she is a sort of a lead cox with the Oxford boat, women's boat team, and spoke quite a bit around that, also around motivation and inspiring people and she certainly won *me* over anyway in her style of communication. She, she, very hard working, very driven, spoke a lot about her sort of study, study style, which passed, I think it was very driven, very balanced as well because she fits in lots of extra curricular activities as well, as well as studying. She had a very good interest in, you know, business and awareness of the accountancy profession, and sort of issues that were impacting us, talked about IFRS and the impact that that might be having impacting on us and our resource requirement which was quite interesting. I mean I thought she was great I thought she, I would say her communication style was very good she was *very* fluent. She did lack a little bit of awareness of when to stop when to wrap up once, she didn't sort of catch on to the – okay time to move on to the, to the next point but I mean I don't see that as a major drawback and I find….but I actually thought she did articulate her points very well, she expanded on them quite a lot but didn't take long to get to the point, she gave a lot of examples I thought she could have stopped, she didn't take long to get to the point but she probably went on and elaborated a bit too much. And she, I mean …very team, I mean very team focused, women's rugby team, hockey team, boating team and she is also a musician and has started a string quartet at Oxford they play at weddings and bar mitzvahs and funerals and the like, and establishing a network there so, I thought she was great.

In the above extract the interviewer begins by expressing a general impression of the candidate "I thought she was excellent" and then repeats this general impression twice again, once in the middle of the feedback, and again at the end. Clearly there is evidence here that the assessor is following aspects of the formal consensus meeting procedure, in that an evaluation against the required dimensions is provided, and a justification for that evaluation is presented also. However, in the process of providing this justification, the assessor also engages in non-prescribed behaviour that is consistent with the operation of informal/latent processes. Not only is the assessor's general impression of the candidate communicated to co-assessors, but information that might lead to the use of stereotypes of the candidate (e.g. that she is Japanese, and a musician who plays in a string quartet) is also communicated. Furthermore, the assessor's overall enthusiasm about the candidate, "one of the best candidates I've interviewed in a long time", can also be construed as actively persuading other assessors that this candidate should be selected rather than merely expressing information about her performance against the dimensions in a detached and objective manner. However, more conclusive evidence of the operation of active persuasion is provided in the following extracts.

Evidence of the Use of Active Persuasion on Assessment Ratings

In the following extract, the second interviewer and the leaderless group discussion assessor have given their feedback on a candidate, but the case study assessor is unavailable. For this reason, the meeting facilitator rightly seeks to move on to a discussion of the next candidate, with the intention of obtaining the case study assessor's feedback later on.

Extract 2

Facilitator: Ok, we'll come back to that one.
Second Interviewer: Definitely an offer, definitely an offer.
Facilitator: Ok we've got to come back to that one.
Second Interviewer r: Well based on what I saw….
Facilitator: Ok.
Second Interviewer: …..I think we'd be very silly not to make her an offer.
Facilitator: Well, we will wait to see what the case study says, but based on what you two have said it looks like it will be an offer, yeah.

This is a graphic example of active persuasion by an assessor. Such a process is also clearly identifiable in the extract below, in which the case study assessor and second interviewer request that the facilitator to adjust the candidate's scores to ensure that he is selected.

Extract 3

Facilitator: Okay, well can we just, quickly say, and we'll jump back to you then, ok? Going across then, we've got two 3's and a 4 just in the end column with everybody, would people be happy with a "good" for that? Yeah? Okay, what about the next column it is two 3's and a 4 again, so it's not great is it?
Second Interviewer: What is the column?
Facilitator: It's practical intellect.
Second Interviewer: Happy with a 3.
Case study assessor: I'm happy, I mean, my general feeling is it we should take him, so whatever…..
Second Interviewer: Yes, whatever it needs to take him, this guy, this guy - we're taking, so….

Facilitator: Yep.
Second Interviewer: ….if you make it fixed so we take him…
General laughter
Facilitator: You want me to fix the end column? Okay, so nobody has got any objections. It looks basically that it's going to be a 4 there.
Second Interviewer: Yeah.
Case study assessor: Yeah.
Facilitator: Okay a 4 there, and 3's for the rest, okay we happy with that?
Second Interviewer: Yeah.
Facilitator: Ok, that's an offer. Brilliant.

Evidence of the Effect of Power on Assessment Ratings: Qualitative Data

My observation of consensus meetings suggested that the use of such power often took the form of differences in how much was said about a candidate, and the force and confidence with which it was expressed. The force and confidence of a contribution is partly manifested in the exact intonation of what is said, and in the non-verbal behaviour of the speaker, and it is not possible to convey this fully in written extracts. However, it is possible to give an example of differences in the nature (and volume) of words used by relatively senior and junior assessors, and the effect this appears to have on the junior assessor's judgment.

In the following extract, the leaderless group discussion (LGD) assessor begins by giving a score of 2 (marginal) for communication skills to a candidate, and this is followed by the second interview who has given the candidate a good score of 4 (strong) on this dimension. A discussion then ensues.

Extract 4

LGD assessor: Practical intellect 3, team work 3, interest in business 4 ... 3 to 4, and communication skills 2.
Facilitator: Alright, okay, take me through the interview first.
Second Interviewer: Again someone who is, is very comfortable in driving to get good results. She is someone who I think is trying to help herself pay through getting though university, so she does a part time job most evenings, goes to University gets some good results, and on top of that also does some voluntary work on the weekends so someone in terms of trying to get good results and commitment, I see no problem there, in there, at all. In terms of responsiveness to change - it was alright – she, she described herself, her biggest issue if you like in terms of change was when she went to university she is someone who grew up in Amersham, has never left Amersham, was very sort of quiet and reserved at school, and, for her the big change was she went to university in a sense it was exposed if you like, and she realised that she needed to do something about that, so the first thing she did was she went and got herself elected to the first year rep for economics which is a pretty bold step actually because for exactly that reasons she needed to, to like to come out of herself which er, certainly in the interview she obviously had, its interesting the comment you made in terms of communication skills in terms of whether she was *really* outward.
LGD Assessor: Hmm.
Second Interviewer: 'Cause certainly in interview she was, but that was one-on-one.
LGD Assessor: Yeah.

Second Interviewer: So, so it would be useful to get some communication skills in the group feedback.
LGD Assessor: Yeah.
Second Interviewer: Because that might demonstrate that actually she is comfortable one-on-one but was less comfortable when she was in a group environment
LGD Assessor: Well I did have, I had 3's scrubbed out and put 2.
Facilitator: Communication is this?
LGD Assessor: Yeah, I thought, I mean she did, really, withdraw from the meeting, em, it was quite a large period where she didn't actually say *anything* at all, certainly towards the end of the last erm…And yeah it was a quiet group, anyway there wasn't any sort of some spark, there was no catalyst there to maybe get people discussing things properly, um, I mean what she said she said it clearly confidently, and so I think I maybe have been a bit harsh on her, erm ….
Facilitator (to interviewer): Would you be happy to go down to a 3?
Second Interviewer: On communication skills? Not in terms of what I was seeing and the examples we had. One-on-one she was very confident, very comfortable, and the example she gave of where she had had to use her communication skills was, you know, as a representative of the first year economics class she had to attend board meetings of the faculty, and present, and she was quiet *comfortable* doing that and that was why she put herself in that situation. So, I mean if that she giving that as a example so…
LGD assessor: I'll happy, I'll happily change mine to a 3 or more.

Here the second interviewer utters approximately four times as many words as the younger and more junior LGD assessor. Furthermore, the senior assessor's speech is clear and confident whereas that of the LGD assessor appears unsure of his evaluation ("Well I did have, I had 3's scrubbed out and put 2"). In his penultimate contribution, the LGD assessor does appear to rally, and provides a sound justification for his low mark, but after interviewer's subsequent firm pronouncement that he is unwilling change his evaluation of the candidate, the LGD assessor capitulates, saying that he is willing to change his mark of 2 (marginal) to 3 (good), or 4 (strong).

This transcript shows how the informal influence of power can operate in ACs, but it does not show the degree to which such influence affects assessment centre outcomes overall. This issue is addressed with the following quantitative analysis of the relative influence of candidates' scores on various assessments (including the second interview) on overall selection decisions.

Evidence of the Effect of Power on Assessment Ratings: Quantitative Data

Of the three types of assessor attending the consensus meetings (i.e. the LGD assessor, the case study assessor, and the second interviewer), the second interviewer was the most senior. Furthermore, because an appointed candidate would work in his or her team, this person had a vested interest in the outcome of meetings. If latent/informal processes involving the use of persuasion and power were influencing the outcome of consensus meetings in the organization being studied, we might expect the assessments of candidates made by the relatively senior, and powerful, second interviewer to have the greatest impact on decisions about whether or not to accept the candidate.

To examine this proposition, data on 413 candidates who had attended the graduate assessment centre in the financial organization were obtained. Overall scores across the eight dimensions were computed for the three assessment centre exercises and the first interview. In the case of the three facets of the cognitive ability test, raw scores were used. Logistic regression analysis was then used to examine the degree to which the performance of candidates on each assessment centre exercise, the first interview, and the cognitive ability test, influenced final selection decisions.

A test of the full model using all predictors against a constant only model was statistically reliable chi-square (7, N=413) = 189.20, p<.0001, indicating that the exercise scores reliably predicted whether or not assessees were offered a post or not. Nagelkerke's R^2 is .57. The model correctly predicted 61% of candidates who were rejected, and 97% of candidates who were accepted. Wald statistics, odds ratios, and the 95% confidence intervals for the odds ratios are shown in Table 1.

Table 1

The Extent to Which the Exercise Scores Predicted Whether or Not 413 Graduate Candidates were Selected

					95% C.I. for odds ratio	
Exercise	**B**	**Wald**	**Sig.**	**Odds ratio**	**Lower**	**Upper**
First Interview	-1.02	5.01	0.0252	0.36	0.15	0.88
Second Interview	3.60	48.45	0.0000	36.49	13.25	100.46
Leaderless group discussion	1.93	27.67	0.0000	6.88	3.35	14.12
Case study	1.45	12.87	0.0003	4.26	1.93	9.40
Verbal ability	0.01	0.54	NS	1.01	0.99	1.02
Numerical ability	-0.01	1.19	NS	0.99	0.97	1.01
Abstract reasoning ability	-0.01	0.64	NS	0.99	0.98	1.01
Constant	-14.97	35.33	0.0000	0.00		

Table 1 shows that, as predicted, the second interview had the largest impact on whether or not candidates were selected: after taking into account the associations with the other predictors, for each additional mark they obtained on the second interview they were 36 times more likely to be selected. In comparison, they were 7 times more likely to be selected for each additional score obtained in the leaderless group discussion, and four times more likely to be selected for each additional score obtained in the case study. These results indicate that the mark given to candidates in the second interview had the greatest unique association with the selection decision. As this second interview was carried out by the most senior assessor, this finding is consistent with the effect of power relationships on selection decisions in ACs.

Interestingly, the sources of information about candidates which had the least influence on the outcome of the meetings, the cognitive ability test results, and the first interview scores, were merely available to assessors, and nobody attending the meetings was made responsible for drawing attention to these potentially useful sources of information. It is worthy of note that despite the well established relation between cognitive ability and job performance (Schmidt & Hunter, 1998), there was no significant relation here between whether the candidates were selected and their cognitive ability test scores.

Discussion

If consensus meetings are construed as short-lived social groups, there is a rich body of research indicating that processes such as normative and informational influence, and the operation of power, will take place, and that these processes are likely to affect decision-making. However, this is the first empirical study designed to examine the existence and nature of several of these informal/latent processes in a consensus meeting.

Most of the literature on ACs focuses on an assessment of their criterion-related validity (Byham, 1970; Cohen et al., 1977; Howard, 1974; Muchinsky, 1986; Thornton & Byham, 1982), and their construct validity (Archambeau, 1979; Neidig & Neidig, 1984; Sackett & Dreher, 1982; Shore et al., 1990), and in doing so researchers normally examine the associations between candidates' scores on the different dimensions and exercises within ACs, and their scores on various measures of job performance (Chan, 1996; Lievens, 2001; Rolland, 1999). The research reported here differs from these themes, and their accompanying research paradigm, in several respects. First, it focuses on assessment centre consensus meetings rather than the assessment centre process more generally. Second, the methodology, rather than being exclusively quantitative, combines qualitative and quantitative approaches. Third, instead of basing the research solely on an examination of the numerical data produced in ACs and making inferences from these, the processes taking place in ACs are studied more directly, using observation and auditory recordings of real ACs. Fourth, rather than viewing the initial candidate ratings processes in ACs, and subsequent consensus meeting discussions, as akin to a sophisticated information processing system, they are construed here as a set of vulnerable cognitive and social processes in which both formal/explicit and informal/latent events take place simultaneously. As a result of these shifts in methodology and emphasis, it has been possible to identify a variety of processes which may interfere with the quality of selection decisions made in ACs.

The auditory recordings of a consensus meeting presented here suggests that, in addition to formal and explicit processes in which assessors have been trained to engage, including evidence-based ratings of candidates against carefully-defined and performance-relevant dimensions, several informal/latent processes can also have an influential impact on the decision making process operating in consensus meetings. First, assessors do not merely bring to a consensus meeting a set of scores, and evidence for those scores, but also an overall impression of the quality of the candidate (Lance et al., 2004; Russell, 1985). Second, they are not merely acting as "information processing machines", or scientists, carefully testing hypotheses by examining evidence from several sources and arriving at a detached and balanced overall view of the suitability of candidates. Whilst they may try to be objective and detached (or at least present themselves as such), it is clear that at least some assessors also use the consensus meeting in general, and the opportunity they are given to provide evidence for their scores in particular, to actively persuade others about the strengths (or weaknesses) of the candidates they have assessed. Furthermore, in the process of trying to persuade other assessors, they may present information about the candidate which is not directly relevant to the dimensions being examined (see Extract 1). Such extraneous information, brings an attendant risk of unwarranted inferences, stereotyping, and bias, and is liable to reinforce the tendency of assessors to adjust their ratings of candidates on the pre-specified dimensions to fit their overall impressions rather than to strictly adhere to evaluations based on observed

behaviours. Finally, transcripts of the auditory recordings of wash-ups, in combination with the logistic regression analysis, suggests that assessors who are perceived as having more power, as a consequence of being more senior within the organization than other assessors, and/or because they are known to be the person who will work with, or manage, a candidate if he or she is appointed, have an inflated influence on the final selection decision. Finally, information that does not have a champion in the consensus meeting (e.g. in the organization studied here, the results of the first interview and the cognitive ability tests) may be underutilized, or even ignored.

The transcripts presented here were taken from a single consensus meeting, and the logistic regression analysis from the use of consensus meetings in a single organization. It could be argued that the latent-informal events identified are actually a set of disturbing outliers rather than examples of events which are common in ACs. There are several reasons to doubt this interpretation. First, the author witnessed clear examples of informal/latent processes in all four consensus meetings he attended. Second, the assessment centre in which the auditory recordings presented here were recorded, took place in a large and very successful multi-national organization in which much emphasis was placed on proper assessor training, and good professional practice in the running of ACs. Third, some of the features of the consensus meetings studied here which are likely to have intensified latent-informal processes, including significant differentials in the seniority of assessors, and the presence of an assessor who would work with an appointed candidate, are probably quite common. Although there is at present no systematic research on differences in the seniority of assessors in consensus meetings, the large scale survey by Spychalski et al. (1997) found that assessors were usually line and staff managers selected somewhat haphazardly, with over 30% used because they volunteered or happened to be available. Such a finding does not suggest that in most consensus meetings assessors are carefully selected to be of equal seniority.

Lastly, although this is the first study to examine the simultaneous presence of the use of persuasion, general impressions of candidates, and the effects of power differentials in consensus meetings, the small amount of previous work on these latent/informal processes in the assessment centre context also detected these phenomenon (Klimoski et al., 1980; Lowry, 1992; Russell, 1985; Sackett & Wilson, 1982), such findings are consistent with the effects of well-established social psychological phenomena such as informational influence, normative influence, and the influence of power (Bishop & Myers, 1974; Deutsch & Gerrard, 1955; French & Raven, 1959; Kaplan & Miller, 1987; O'Reilly & Caldwell, 1979; Raven, 1965, 1993; Turner et al., 1989; Wood et al., 1994).

Practitioner Implications

If we consider (a) the evidence that ACs which exclude consensus meetings have the same, or greater, criterion-related validity than those which include them (Feltham, 1988; Herriot, 2003; Jones et al., 1991; Pynes & Bernardin, 1989), (b) the manifest psychological and social psychological events and processes identified in the research presented in this article, (c) the inevitable difficulties associated with monitoring and averting the operation of these events and processes, some of which, like the influence of general impressions of candidates on assessor feedback, may be non-observable, and (d) the significant costs associated with running consensus meetings, there seems little or no justification for their continued use. Instead,

information about candidates can be combined in a variety of arithmetical ways, from simply summing the scores obtained by a candidate and examining whether it meets some pre-assigned threshold, to the use of sophisticated weighted-averaging systems developed by regressing measures of job performance on the results of assessment centre exercises.

However, it is possible that within organizations the suggestion that consensus meetings should be abolished would be met with resistance. With them, it is possible for organizational gate-keepers to use their influence in these meetings to exclude candidates who they feel have poor organizational-fit (Herriot, 2003), disguising the impact of their overall impression of the candidate with articulate, persuasive, and seemingly detached and rationale rhetoric couched, at least partially, as feedback against the chosen dimensions. Indeed, in extreme cases, the process of integrating and evaluating scores carefully obtained by trained assessors may become little more than a stage for political manoeuvring.

This raises issues to which organizational psychologists have hitherto perhaps paid insufficient attention. The process of forming general impressions of job candidates, making a decision about whether they should be appointed, and then using consensus meetings to persuade other assessors that they should concur with this decision, is essentially the operation of organizational politics. Few authors, with the notable exception of Ferris and King (1991), have addressed the politics of personnel selection, yet the powerful effects of this process may arise where they are least expected. The assessment centre, with its emphasis on obtaining information about candidates on clearly defined and job-relevant dimensions, and doing so using multiple trained assessors and multiple methods, and on combining this information through rational discussion, represents, at least in appearance, the best exemplar of a rational and fair model of personal selection. And for this very reason it provides an ideal cloak with which to conceal the operation of subtle but powerful forms of bias and discrimination.

Such bias and discrimination does not necessarily reflect conscious or unconscious attempts to exclude the members of obvious target groups, such as people from ethnic minorities, women, or disabled people. Rather it works to ensure that those who are selected have the right "fit" in that they are "the 'right types' who reflect the proper 'chemistry' and thus fit in well with the organizational environment and culture" (Ferris & King, 1991, p.62). Theory and research in personnel selection has tended to place emphasis on providing practitioners with the best possible technology to identify people who are likely to be the best job performers. But, as Herriot points out, people who manage teams and departments are likely to want people who are not only "good", but who will also "fit in". Unless those involved in selection research confront this issue directly, and try to find ways of enabling practitioners to obtain good person-organization fit, or good person-team fit, without engaging in unjustified discrimination (Latham, 1995; Ryan & Schmit, 1992), it is likely that practitioners will continue to use covert political steps to circumvent the unwanted outcomes of apparently rational selection processes, including ACs and the consensus meetings which usually attend them.

References

Archambeau, D. (1979). Relationships among skill ratings assigned in an assessment center. *Journal of Assessment Center Technology, 2*(7-20).

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*(1), 125-154.

Bender, J. M. (1973). What is "typical" of assessment centers. *Personnel, 50*, 50-57.

Bishop, G. D., & Myers, D. G. (1974). Informational influence in group discussion. *Organizational Behavior and Human Performance, 12*(1), 92-104.

Boyle, S., Fullerton, J., & Wood, R. (1995). Do assessment/development centres use optimum evaluation procedures? A survey of practice in uk organizations. *International Journal of Selection and Assessment, 3*(2), 132-140.

Byham, W. C. (1970). Assessment center for spotting future managers. *Harvard Business Review, 48*, 150-160.

Caldwell, C., Thornton, G. C., & Gruys, M. L. (2003). Ten classic assessment center errors: Challenges to selection validity. *Public Personnel Management, 32*(1), 73-88.

Cascio, W. F. (1986). *Managing human resources*. New York: McGraw Hill.

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology, 69*, 167-181.

Cohen, B. M., Moses, J. L., & Byham, W. C. (1977). *The validity of assessment centers: A literature review*. Pittsbugh, PA: Development Dimensions Press.

Deutsch, M., & Gerrard, H. B. (1955). A study of normative and informational social influence upon individual judgement. *Journal of Abnormal and Social Psychology, 51*(629-636).

Feltham, R. (1988). Assessment-center decision-making - judgemental vs mechanical. *Journal of Occupational Psychology, 61*(3), 237-241.

Ferris, G. R., & King, T. R. (1991). Politics in human-resources decisions - a walk on the dark side. *Organizational Dynamics, 20*(2), 59-71.

French, J. R. J., & Raven, B. H. (1959). The bases of social power. In D. Cartwright (Ed.), *Studies in social psychology* (pp. 150-167). Ann Arbour, MI: Institute for Social Research.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*(3), 493-511.

Gilbert, P. J. (1981). An investigation of clinical and mechanical combination of assessment center data. *Journal of Assessment Center Technology, 4*, 1-10.

Herriot, P. (2003). Assessment by groups: Can value be added? *European Journal of Work and Organizational Psychology, 12*(2), 131-145.

Highhouse, S., & Harris, M. M. (1993). The measurement of assessment-center situations - bem template matching technique for examining exercise similarity. *Journal of Applied Social Psychology, 23*(2), 140-155.

Howard, A. (1974). An assessment of assessment centers. *Academy of Management Journal, 17*, 115-134.

Hunter, J. E. (1980). Test validation for 12000 jobs: An application of synthetic validation and valdidity generalizaton to the general aptitude test battery (gatb): Michigan State University.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Jones, A., Herriot, P., Long, B., & Drakeley, R. (1991). Attempting to improve the validity of a well-established assessment-center. *Journal of Occupational Psychology, 64*(1), 1-21.

Kaplan, M. F., & Miller, C. E. (1987). Group decision-making and normative versus informational influence - effects of type of issue and assigned decision rule. *Journal Of Personality And Social Psychology, 53*(2), 306-313.

Keenan, T. (1995). Graduate recruitment in britain - a survey of selection methods used by organizations. *Journal of Organizational Behavior, 16*(4), 303-317.

Klimoski, R. J., Friedman, B., & Weldon, E. (1980). Leader influence in the assessment of performance. *Personnel Psychology, 33*, 389-401.

Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal Of Applied Psychology, 89*(1), 22-35.

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13*(4), 323-353.

Latham, G. P. (1995). Using the situational interview to assess organizational fit. *Canadian Psychology-Psychologie Canadienne, 36*(2A), 8-8.

Lievens, F. (2001). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior, 22*(3), 203-221.

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment centre process: Where are we now? In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (Vol. 16). New York: Wiley.

Lowry, P. E. (1992). The assessment center: Effects of varying procedures. *Public Personnel Management, 21*(2), 171-183.

Lowry, P. E. (1997). The assessment center process: New directions. *Journal of Social Behavior and Personality, 12*(5), 53-62.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews - a comprehensive review and metaanalysis. *Journal of Applied Psychology, 79*(4), 599-616.

Muchinsky, P. M. (1986). Personnel selection methods. In C. Cooper & I. Robertson (Eds.), *International review of industrial and organizational psychology*. New York: Wiley.

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment-center exercises and job relatedness. *Journal of Applied Psychology, 69*(1), 182-186.

O'Reilly, C. A., & Caldwell, D. F. (1979). Informational influence as a determinant of perceived task characteristics and job-satisfaction. *Journal of Applied Psychology, 64*(2), 157-165.

Pynes, J. E., & Bernardin, H. J. (1989). Predictive-validity of an entry-level police officer assessment-center. *Journal of Applied Psychology, 74*(5), 831-833.

Raven, B. H. (1965). Social influence and power. In I. D. Steiner & M. Fishbein (Eds.), *Current studies in social psychology* (pp. 371-381). New York: Holt, Rinehart & Winston.

Raven, B. H. (1993). The bases of social power: Origins and recent developments. *Journal of Social Issues, 49*(227-252).

Rolland, J. P. (1999). Construct validity of in-basket dimensions. *European Review of Applied Psychology-Revue Europeenne De Psychologie Appliquee, 49*(3), 251-259.

Russell, C. J. (1985). Individual decision-processes in an assessment center. *Journal of Applied Psychology, 70*(4), 737-746.
Ryan, A. M., & Schmit, M. J. (1992). Validation of an organizational fit instrument. *International Journal Of Psychology, 27*(3-4), 509-509.
Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment-center dimensions - some troubling empirical-findings. *Journal of Applied Psychology, 67*(4), 401-410.
Sackett, P. R., & Wilson, M. A. (1982). Factors affecting the consensus judgment process in managerial assessment-centers. *Journal of Applied Psychology, 67*(1), 10-17.
Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.
Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment-center dimension and exercise constructs. *Journal of Applied Psychology, 77*(1), 32-41.
Shore, T. H., Thornton, G. C., & Shore, L. M. (1990). Construct-validity of 2 categories of assessment-center dimension ratings. *Personnel Psychology, 43*(1), 101-116.
Spychalski, A. C., Quińones, M. A., Gaugler, B. B., & Pohley, K. (1997). A survey of assessment centre practices in organizations in the united states. *Personnel Psychology, 50*, 71-90.
Tajfel, H., & Turner, J. C. (1986). The social identiy theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations*. Chicago: Nelson-Hall.
Thornton, G. C. I., & Byham, W. C. (1982). *Assessment centers and managerial performance*. New York: Academic Press.
Turner, J. C. (1986). Social categorization and the self-concept: A social-cognitive theory of group behavior. In E. J. Lawler (Ed.), *Advances in group processes: Theory and research* (Vol. 2). Greenwich, C.T.: JAI press.
Turner, J. C., Wetherell, M. S., & Hogg, M. A. (1989). Referent informational influence and group polarization. *British Journal of Social Psychology, 28*, 135-147.
Wood, W., Lundgren, S., Ouellette, J. A., Busceme, S., & Blackstone, T. (1994). Minority influence: A meta-analytic review of social influence processes. *Psychological Bulletin, 115*, 323-345.