Assessment Center Procedures: Cognitive Load During the Observation Phase

Nanja J. Kolk,
Berenschot Business Management, Utrecht
Henk van der Flier
Free University, Amsterdam
Marise Ph. Born,
Erasmus University, Rotterdam
Juliette M. Olman*
Berenschot

This study explores the traditional procedure of observing assessment center exercises while taking notes vs. an alternative procedure where assessors merely observe and postpone note-taking until immediately after the exercise. The first procedure is considered to be cognitively demanding due to the requirement of simultaneous note-taking and observing. Also, dual task processing (concurrent observing and note-taking) is considered to be especially demanding for assessors without rating experience. The procedures are evaluated using a 2 × 2 design (with note-taking/without note-taking × experienced/inexperienced). Some 121 experienced and inexperienced assessors rated videotaped candidates, observing either with or without taking notes. Results showed that experienced assessors yield significantly higher differential accuracy than inexperienced assessors. We did not find an effect of observation procedure on accuracy, interrater reliability or halo. Implications for future research are described.

Introduction

It is well known that observing candidates in an assessment centre (AC) is a highly cognitively demanding task for the assessors (e.g., Gaugler and Thornton 1989; Zedeck 1986). The traditional AT&T observation procedure, applied in the majority (80–95%) of the ACs (Spychalski, Quiñones, Gaugler and Pohley 1997; Thornton 1992), involves four independent phases: during an exercise assessors observe a candidate's behaviour and take notes of these observations at the same time. Afterwards, assessors classify the written remarks into behavioural dimensions, then, they give a quantitative rating per dimension. Finally, they evaluate their ratings with co-assessors (Bray, Campbell and Grant 1974).

The first phase in this procedure involves two tasks: observing and note-taking. Note-taking during the exercise is considered an essential part of the rating process (Task Force on Assessment Center Guidelines 1989). It presumably helps the assessor to memorize the observed behaviours (Thornton 1992). The potential downside is, and this is this study's focal point, that note-taking forces the assessor to perform the observation and recording tasks simultaneously. This dual task may require more cognitive processing capacity than assessors have available, potentially leading to more rating errors. Also, the assessors' attention has to shift from the exercise to their record-sheet, so that key behavioural information might be overlooked (Hennessey, Mabey and Warr 1998). Taking notes during observation may therefore be a mixed blessing.

A second and more practical incentive for conducting this study is that, in practice, the writing of the behavioural report seems to occur rather unsystematically. Some assessors record few observations, some record a lot. Also, the recording of the behaviours is not always maintained until the end of the exercise (Hennessey *et al*. 1998, p. 229; Lievens and Goemaere 1999). In addition, assessors are often not only passive observers but also active participants in an exercise, such that they play the role of the subordinate (Zedeck 1986). This active role-playing obviously prohibits taking notes at the same time. Is this lack of systemization problematic? This study begins to explore the question of whether note-taking is cognitively

* Address for correspondence: Nanja Kolk, Berenschot, Europalaan 40, 3526 KS Utrecht, The Netherlands. E-mail: N.Kolk@Berenschot.com

demanding, due to the requirement of processing information while performing two tasks concurrently. We study the effect of postponing note-taking until immediately after the exercise has ended.

There is not a large body of empirical research on this matter. Zedeck (1986, p. 272) states that few, if any, of the theories and results in cognitive psychology and information-processing literature have been applied to the AC method. Nonetheless, the basic cognitive model for evaluation of performance may, in general terms, be applicable to ACs as well. This model involves several phases (Cooper 1981; Murphy and Cleveland 1995; Zedeck 1986). Translated to the AC context, these phases are as follows: first, during each exercise the assessors *observe* a candidate's behaviour, paying special attention to those behaviours that relate to the target dimensions. Then, the assessor *encodes* the information about the behaviour. This involves a complex process of *categorization* that may occur automatically (automatic categorization) but often requires careful attention and mental effort (controlled categorization), depending on the complexity of the information to be processed (Murphy and Cleveland 1995, p. 189). In the context of the AC, it is likely that most of the complex managerial behaviours are processed in the controlled mode. Assessors need to determine actively whether or not a behaviour is a member of a dimension category (Zedeck 1986). Subsequently, the behavioural information is *stored in memory* for later use, which is mostly immediately after the exercise. When making the rating, the assessor must *retrieve* information from memory and *integrate* all relevant bits of information on the candidate.

By writing down observations during the exercise, the material is coded verbally, which facilitates storing the observations in memory. However, the dual task requirement may also increase cognitive load. Controlled processing of information, which occurs for novel, unusual, and complex observations, such as observations of managerial behaviour in AC exercises, is highly demanding on processing capacity. This makes it difficult to perform several tasks at the same time (Murphy and Cleveland 1995, p. 192). Consequently, assessors may make more observational and rating errors, such as failing to notice key behaviours, incorrectly appointing behaviours to dimensions, etc., leading in the end to lower inter-rater reliability, accuracy and construct validity. Cognitive load has been posited to be largely responsible for rater errors in AC ratings (e.g., Zedeck 1986). AC literature supports this notion, in that experienced assessors produce more valid ratings than inexperienced assessors (Gaugler, Rosenthal, Thornton and Bentson 1987; Sagie and Magnezy 1997) and are less apt to contrast effects between good and poor performing candidates (Gaugler and Rudolph 1992). This may be explained by the finding that practice with a given task leads to a decrease in the necessary cognitive resources (Best 1992). Naturally, the more complex a task is, the more practice will be needed before this task can be adequately performed together with another task. Thus, inexperienced assessors, such as managers who act as assessors infrequently, may lack the necessary practice to be able to perform two tasks concurrently. It is essential for this group of assessors that the observation phase is as little cognitively demanding as possible. Dual task processing may be especially demanding for assessors without rating experience.

This study explores the effects of concurrent note-taking and observing during an AC exercise. We aim to contribute to the limited body of literature that exists on this topic by examining two alternative observation methods: the traditional method, which requires note-taking and observing during the exercise, and an alternative observation method, in which assessors are merely required to observe without taking notes. In the latter method, assessors observe behaviours during each exercise and postpone writing down the observations until immediately after the exercise. Because the assessors produce a behavioural report the moment the exercise has ended, they avoid the risk that the level of detail they recall will be negatively affected, as suggested by Cooper (1981). We compare the two observation methods here in terms of inter-rater reliability, halo and accuracy.

The answer to the question whether note-taking during an exercise is too cognitively demanding may be sought in an interaction effect between observation method and assessor type. First, we hypothesize that the traditional method is more appropriate for experienced assessors than the 'observe only' method. This is expected because, after practice, both the observation and the writing task are (at least partly) automated, thus facilitating the dual task demand. Not being allowed to take notes may increase cognitive demands in the 'observe only' method, because observations need to be stored in memory. Also, not taking notes may seem unnatural to experienced assessors, because they are used to being able to write down whatever they perceive as important information. The traditional method is hypothesized to be more cognitively demanding for inexperienced assessors than the 'observe only' method, due to concurrent observation and writing.

Based on previous research, we also hypothesize that overall, experienced assessors yield higher accuracy than inexperienced assessors.

Method

Summary

Experienced and inexperienced assessors evaluated the performance of three candidates taped on video. The candidates participated in two interview simulations.

Both groups of assessors received a traditional training, in which the behavioural dimensions were explained and rater errors were described. In the training session, the assessors rating according to the traditional method were asked to make a behavioural report during observation, whereas the assessors rating according to the 'observe only' method were asked not to make a behavioural report during the exercise, but to postpone writing down relevant observations until immediately after the exercise.

Participants

Participants in this study were 121 assessors (34 were male). Within this group, 31 assessors had rating experience (i.e., mean years of rating experience = 5.8 years, $SD = 6.6$). Experienced assessors were mostly HR specialists with an MA in Psychology, who serve as assessors on a day-to-day basis. The inexperienced assessors were Master's students, mostly majoring in Work and Organizational Psychology. Thus, the inexperienced and experienced assessors had a very similar educational background, although none of the inexperienced had previously worked as an assessor. The use of this type of students seems appropriate for the present research, since W/O Psychology students will often be asked to serve as assessors in their future job as an HR specialist. The experienced assessors volunteered to participate in the current study. The students received a small fee for their participation. Both groups of assessors were randomly assigned one of the two observation methods.

Assessor Training

The experienced assessors had already received recurring assessor training sessions, focusing on the meaning of the dimensions, on rating errors and on a shared frame of reference. Because the materials used in this experiment were obtained from the working environment of the experienced assessors, there was no need to spend time explaining the dimension definitions. Therefore, specific instructions for experienced assessors prior to participating in this experiment focused mainly on the observation procedure.

Inexperienced assessors received a training session to familiarize them with the basic features and techniques of the AC. The instructor gave them a detailed description of the task requirements of an assessor, the assessment procedure (observing, classifying, rating and evaluating) and common rater errors (halo, leniency/severity). In addition, in-depth descriptions of the dimension definitions were given, including behavioural descriptions and examples of effective and ineffective behaviours per dimension for each exercise they were about to observe. Assessors were also presented with the role-instructions for both the candidate and the subordinate. They were given the opportunity to ask questions and receive feedback. The training given to the inexperienced assessors was similar in content to the training that had been given to the experienced assessors.

Subsequently, both groups of assessors were either instructed to take notes during the exercise (traditional method), or to postpone note-taking until immediately after the exercise was completed ('observe only' method).

Simulated Assessment Center

The simulated AC consisted of six videotaped interview simulations (three candidates participating in two different interview simulations). Research has shown that using videotaped rather than 'real-life' simulations does not affect accuracy of the ratings (Ryan, Daum, Bauman, Grisez, Mattimore, Nalodka and McCormick 1995). In the first exercise, candidates were to win over the reluctance of a rather self-conscious and busy subordinate to give a lecture for a large audience, the day after tomorrow. In the second exercise, the candidates were required to persuade a subordinate to put in overtime on a Friday afternoon, which the subordinate was initially unwilling to do. The roles of the subordinates were played by six trained role-players. Performance dimensions were sensitivity, co-operation, judgment, tenacity and effectiveness. The fidelity of these AC simulations was ascertained because the stimulus materials that were used in the videos (role instructions, dimension definitions, rating forms, etc.) were obtained from an operational AC.

Evaluation Criteria

Accuracy was measured by comparing observed assessor ratings to an estimated 'true score' (Table 1). This true score was obtained by a panel of three experts (i.e., highly experienced) assessors (two men and one woman; mean age = 34 years; mean rating experience = 6.5 years). These experts were asked to rate each candidate in each exercise, under optimal conditions (Murphy and Cleveland 1995, p. 290). That is, they were allowed to take as much time as needed, to rewind the video-tapes, and to discuss the candidates' performance until they reached consensus for each dimension (see Table 1 for the rating profiles of each candidate). The difference between this optimal rating procedure and the normal rating procedure followed by the 121 participating assessors, is that the latter group was allowed to view the videotapes only once, had only 15 minutes to make their rating, and were asked not to discuss their findings with co-assessors. Expert ratings are generally regarded as a good measure of true performance (Gaugler and Rudolph 1992; Jones 1997; Smither, Barry and Reilly 1989).

Table 1. Candidate 'true scores' obtained by expert ratings

	Candidate Profile		
	Candidate 1	Candidate 2	Candidate 3
Exercise 1			
Sensitivity	2.5	2.5	3
Judgement	3	3.5	3
Tenacity	4	4	3
Effectiveness	2.5	3	3
Exercise 2			
Co-operation	3	1.5	2
Judgement	2	2.5	2.5
Tenacity	2	4	3.5
Effectiveness	2	1.5	2

Note: Rating scale ranging from 1 (low score) to 5 (high score).

Cronbach (1955) distinguished two types of accuracy: differential elevation (DE), and differential accuracy (DA). Differential elevation refers to the accuracy in discriminating among candidates, averaging over dimensions. This pertains to how (un)favourably assessors judge candidates compared to how (un)favourably they are rated by experts, across dimensions. DE implies accuracy on an overall assessment rating level. Differential accuracy refers to accuracy in detecting candidate differences in patterns of performance. This implies accuracy in distinguishing both among candidates and dimensions. This pertains to how (un)favourably assessors judge candidates compared to how (un)favourably they are rated by experts, on a dimension level.

Given that the AC's main interest is a candidate's performance per dimension, DA is largely advocated to be the most informative and theoretically relevant outcome (e.g., Lievens 2001; Schleicher and Day 1998). On the other hand, others argue that DE is as important, if not more important than DA. DA is argued to be relevant only when different decisions are being made depending on the pattern of performance (Murphy and Cleveland 1995, p. 287). In the case of ACs, the dimensions that are being measured are selected on the basis of job analysis. Therefore, candidates are to score high on all dimensions in order to get selected, because all dimensions have been considered as relevant. Thus, one might argue that there is no diversified pattern of AC performance in reference to the selection decision. In this light, DE and DA are both relevant to this area of research. Therefore, this study compares the two observation methods and assessor types in terms of DE and DA.

The results showed that DE and DA are uncorrelated ($r = .02$, $p = .80$). This result is commonly reported in the literature (Murphy and Cleveland 1995). We applied multiple analysis of variance on assessor type and observation method, with the two types of accuracy indices as the dependent variables, followed by two separate univariate analyses.

Inter-rater reliability was studied by computing intra-class coefficients (ICC) per dimension, exercise and candidate. These coefficients were then averaged, yielding an overall inter-rater reliability per observation method and per assessor type.

Halo refers to the phenomenon that a rater's general impression of candidates distorts their perception on specific dimensions. Following Hennessey *et al.* (1998), we also examined correlations between dimensions within exercises for the two methods and types of assessors. These correlations can be regarded as measures of halo, although their meaning differs from those in studies using a regular assessee × dimension design, instead of an assessor × dimension design. We applied Fisher's Z test to examine statistical significance of the differences between the Pearson product moment correlation coefficients.

Results

Accuracy

The lower the accuracy score, the more closely the rating of an assessor matches the rating of the experts. The multivariate test revealed a non-significant assessor type × method interaction (Pillai's trace = .02; $F[2, 115] = 1.12$; $p > .05$). This test showed a significant main effect of assessor type (Pillai's trace = .08; $F[2, 115] = 5.13$, $p < .01$) but not for observation method (Pillai's trace = .00; $F[2, 115] = .12$, $p > .05$). Because of these results, we looked at the univariate analyses (Table 2) to discover which relationship assessor type had with the respective

Table 2. MANOVA on DE and DA with assessor experience, observation method and the interaction term of assessor experience × method

Source	Dependent variable	Type III SS	df	MS	F	ES[a]
Assessor type (AT)	DE	.06	1	.06	2.90	.12
	DA	.03	1	.03	6.53*	.21
Observation method (OM)	DE	.00	1	.00	.14	–
	DA	.00	1	.00	.00	–
AT × OM	DE	.04	1	.04	1.93	.09
	DA	.00	1	.00	.50	–
Error	DE	2.30	116	.02		
	DA	.53	116	.00		
Corrected total	DE	2.45	119			
	DA	.57	119			

Notes: DE: differential elevation; DA: differential accuracy; [a]ES: population effect sizes were estimated by taking the square root of (F−1)/(F+N−1), see for instance Hays (1972, p. 327).
* $p < .05$.

accuracy measures. Table 2 shows that experienced assessors outperformed inexperienced assessors in DA and DE, but that the difference on DE was not significant. Effect sizes are .21 and .12, respectively. This is displayed visually in the profile plots for the two indices in Figure 1. Figure 1 implies that inexperienced assessors perform better on DE in the 'observe only' method than in the traditional method. We determined whether this improvement was significant within the inexperienced group. This turned out not to be the case ($F[1, 87] = 3.30$, $p = .07$, effect size = .16). Note that these are smaller groups which decreased the power of this test.

Inter-Rater Reliability

Table 3 shows the intra-class correlations (ICC) for observation methods. The ICCs are higher for the 'observe only' method than for the traditional method (mean $r = .93$ vs $r = .85$). This difference is mostly due to an unusually low ICC for sensitivity in the traditional method. Table 4 offers the ICCs for assessor types. No significant differences appeared between experienced and inexperienced assessors.

We tested whether the coefficients in Tables 3 and 4 were significantly different, using the Hakstian and Whalen (1976, p. 224) procedure for equal *n*s (number of candidates) and unequal Js (in this case, number of assessors). None of the ICCs differed significantly between observation methods or types of assessors. This result may, however, have been influenced by the fact that we used only three videotaped candidates, resulting in low power for the test.

Halo

Table 5 shows mean inter-correlations between dimensions within exercises. Fisher's *Z* tests revealed no significant differences between observation methods and type of assessor.

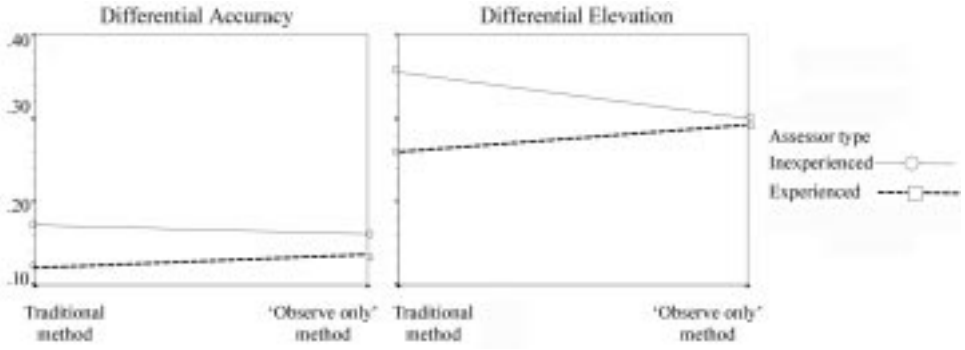

Figure 1. Profile plots for differential accuracy and differential elevation for observation method and exercise type. The lower the score, the higher the accuracy

Table 3. Intra-class coefficients for observation method

Exercise 1	Exercise 2
Traditional method ($N = 62$)	
Sensitivity	.53
Co-operation	.92
Judgement	.90
Judgement	.87
Tenacity	.93
Tenacity	.93
Effectiveness	.80
Effectiveness	.88
Overall mean	.85
'Observe only' method ($N = 59$)	
Sensitivity	.91
Co-operation	.90
Judgement	.88
Judgement	.93
Tenacity	.99
Tenacity	.97
Effectiveness	.88
Effectiveness	.97
Overall mean	.93

Table 4. Intra-class correlation coefficients for assessor type

Exercise 1	Exercise 2
Inexperienced assessors ($N = 90$)	
Sensitivity	.84
Co-operation	.94
Judgement	.94
Judgement	.94
Tenacity	.96
Tenacity	.97
Effectiveness	.92
Effectiveness	.96
Overall mean	.93
Experienced assessors ($N = 31$)	
Sensitivity	.96
Co-operation	.90
Judgement	.97
Judgement	.86
Tenacity	.64
Tenacity	.93
Effectiveness	.94
Effectiveness	.93
Overall Mean	.89

Discussion

This study explores the traditional method of observing AC exercises, which requires simultaneous observing and note-taking, and the 'observe only' method, in which the note-taking part is postponed until immediately after the exercise. Using a 2 × 2 design, these two methods are compared for their suitability for experienced and inexperienced assessors, in terms of two indices of accuracy (differential elevation and differential accuracy), inter-rater reliability and halo.

Overall, experienced assessors outperform inexperienced assessors in terms of differential accuracy (DA) and differential elevation (DE), but for differential elevation the difference is not significant. There appear to be no significant effects of observation method, nor is there a significant observation method × assessor type interaction effect. The results only weakly support our

Table 5. Mean Pearson Product Moment correlation coefficients between dimensions within exercises

		Candidate 1	Candidate 2	Candidate 3
Traditional method	Exercise 1	0.31	0.53	0.64
	Exercise 2	0.43	0.32	0.52
	Mean	.46		
'Observe only' method	Exercise 1	0.26	0.54	0.67
	Exercise 2	0.29	0.45	0.51
	Mean	.45		
Experienced assessors	Exercise 1	0.22	0.73	0.59
	Exercise 2	0.29	0.38	0.66
	Mean	.48		
Inexperienced assessors	Exercise 1	0.27	0.40	0.62
	Exercise 2	0.37	0.35	0.41
	Mean	.40		

expectation that, in terms of accuracy, the 'observe only' method is more suitable for inexperienced assessors and the traditional method suits the experienced assessors better.

The result that experienced assessors yield significantly higher DA than inexperienced assessors confirms our expectations. Apparently, practice with the rating task increases accuracy in detecting patterns within a candidate's performance. The result that experienced assessors do not score significantly higher on DE than inexperienced assessors may indicate that, whereas assessor experience leads to the ability to disentangle behaviours per dimension (DA), assessors differ less with respect to the ability to distinguish bad candidates from good ones (DE). However, it is important to note that the small sample size of the group of experienced assessors may have decreased the power of this test. Conclusions can only be drawn with reservation.

Overall, the 'observe only' method yields somewhat higher inter-rater reliability (.93) than the traditional method (.85), yet this difference is not significant. The low number of candidates ($n = 3$) may have hindered reaching a level of significance. Still, the overall pattern of ICCs between observation method does not deviate substantially. There is also no significant difference between assessor groups in terms of inter-rater reliability.

As regards to halo, the presumed decrease in cognitive load in the 'observe only' method for inexperienced assessors was expected to lower the dimension intercorrelation. Also, previous research has shown experience with the rating task to reduce halo (e.g., Gaugler *et al*. 1987; Sagie and Magnezy 1997). However, results showed no significant differences between observation methods or types of assessors. This result is similar to Hennessey *et al*. (1998) who also failed to find decreased inter-dimension correlation when they compared the traditional observation method with the behavioural checklists and coding method (this latter method involved tallying behaviours during the exercise on a behavioural checklist, rather than recording them). Yet, it is important to realize that both this study and the Hennessey *et al*. study applied a design divergent from the regular candidate × dimension designs to study halo, because this design examines the ratings of multiple assessors and only a few candidates. This assessor × dimension matrix reveals information concerning assessor properties, rather than properties of the candidates. Another explanation is that the candidate profiles used in the present study were too flat (not variable enough) to be able to adequately be distinguished in either of the two observation procedures.

We acknowledge several limitations of the present study. First, the limited sample size of the experienced group of assessors and, second, the small number of candidates taped on video, may have inhibited reaching a level of significance. Another limitation is that, besides differing in rating experience, the two assessor types may have differed in other kinds of experience (e.g., work experience), which may have affected the results.

Study Implications

Although the present results and the results by Hennessey *et al*. (1998) do not support that taking notes increases the quality of the ratings, it cannot be uniformly concluded that assessors might as well skip taking notes during the exercise. The Guidelines (Task Force on Assessment Center Guidelines 1989) specifically state that a systematic procedure must be used by assessors to record accurately specific behavioural observations at the time of their occurrence. Practice seems to go along with this requirement, as none of the surveyed respondents in the Boyle, Fullerton, and Wood (1998) study indicated that no record was made at the time of observation. Yet, this study and the Hennessey *et al*. study do not confirm the importance of this procedure. Also, Hennessey *et al*. found that the quality of the behavioural report did not decline when no written report was produced. Hence, the magnitude of the effect when no record is made may have been overestimated. Along similar lines, the finding that practitioners lack systemization in recording observations during AC exercises seems not as problematic as we assumed at first.

Previous work showed that psychologists produced better discriminant and predictive validity than managers (Gaugler *et al*. 1987; Sagie and Magnezy 1997). Tziner, Ronen, and Hacohen (1993) found that the predictive validity of psychologists' vs job experts' ratings depends on the type of criterion. In their study, the two types of assessors did not differ in predicting job potential, yet job experts were better in predicting job performance. Most recently, it was shown that the superiority of the validity of psychologists' ratings vs job experts' ratings only holds for the prediction of interpersonal style dimensions, and not for performance style dimensions (Damitz, Manzey, Kleinmann and Severin 2000). Recently, Lievens (2001) showed that managers were more accurate assessors than psychology students. The present study adds to these results by showing that a degree in psychology does not warrant the ability to accurately evaluate the performance of applicants on a dimension level, and that experience with the rating task is needed. It seems that two characteristics are important in this respect: the assessors' background (managers evaluate on the basis of their prior knowledge of what is effective and ineffective behaviour) and the assessors' previous rating experience (generally, the more rating experience, the more accurate). Thus, an implication for practice that does follow from this and previous studies is that all assessors – managers and psychologists – alike need thorough rating experience in order to be able to accurately differentiate among dimensions.

Future studies may be directed at further examining the source of cognitive load for assessors. 'Think aloud' studies may focus on capturing what tends to be remembered or forgotten when the assessor does not take notes. Future research may also examine the quality of the behavioural report when it is written either during or after observing the AC exercise.

Acknowledgements

This research was supported in part by LTP, Amsterdam. We also acknowledge Peter Dekker and Dick Neeleman for their comments on the analyses.

References

Best, J.B. (1992) *Cognitive Psychology*, 3rd edn. St Paul, MN: West Publishing Company.

Boyle, S., Fullerton, J. and Wood, R. (1998) Do assessment/development centres use optimum evaluation procedures? A survey of practice in UK organizations. *International Journal of Selection and Assessment*, **3**, 132–140.

Bray, D.W., Campbell, R.J. and Grant, D.L. (1974) *Formative Years in Business: A Long-Term AT&T Study of Managerial Lives*. New York: Wiley.

Cooper, W.H. (1981) Ubiquitous halo. *Psychological Bulletin*, **90**, 218–244.

Cronbach, L.J. (1955) Processes affecting scores on 'understanding of others' and 'assumed similarity'. *Psychological Bulletin*, **52**, 177–193.

Damitz, M., Manzey, D., Kleinmann, M. and Severin, K. (2000) Assessment center for pilot selection: Construct and criterion validity and the impact of assessor type. Manuscript submitted for publication.

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

Gaugler, B.B. and Rudolph, A.S. (1992) The influence of assessee performance variation on assessors' judgments. *Personnel Psychology*, **45**, 77–98.

Gaugler, B.B. and Thornton, G.C. (1989) Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology*, **74**, 611–618.

Hakstian, A.R. and Whalen, T.E. (1976) A k-sample significance test for independent alpha coefficients. *Psychometrika*, **41**, 219–231.

Hays, W.L. (1972) *Statistics*. New York: Holt, Rinehart and Winston, Inc.

Hennessey, J., Mabey, B. and Warr, P. (1998) Assessment centre observation procedures: An experimental comparison of traditional, checklist and coding methods. *International Journal of Selection and Assessment*, **6**, 222–231.

Jones, R.G. (1997) A person perception explanation for validation evidence from assessment centers. *Journal of Social Behavior and Personality*, **12**, 169–178.

Lievens, F. (2001) Assessor training strategies and their effects on accuracy, inter-rater reliability and discriminant validity. *Journal of Applied Psychology*, **86**, 255–264.

Lievens, F. and Goemaere, H. (1999) A different look at assessment centres: Views of assessment centre users. *International Journal of Selection and Assessment*, **7**, 215–219.

Murphy, K.R. and Cleveland, J.N. (1995) *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives*. Thousand Oaks, CA: Sage.

Ryan, A.M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T. and McCormick, S. (1995) Direct, indirect and controlled observation and rating accuracy. *Journal of Applied Psychology*, **80**, 664–670.

Sagie, A. and Magnezy, R. (1997) Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology*, **70**, 103–108.

Schleicher, D.J. and Day, D.V. (1998) A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behaviour and Human Decision Processes*, **73**, 76–101.

Smither, J.W., Barry, S.R. and Reilly, R.R. (1989) An investigation of the validity of expert true score estimates in appraisal research. *Journal of Applied Psychology*, **74**, 143–151.

Spychalski, A.C., Quiñones, M.A., Gaugler, B.B. and Pohley, K. (1997) A survey of assessment center practices in organizations in the United States. *Personnel Psychology*, **50**, 71–90.

Task Force on Assessment Center Guidelines (1989) Guidelines and ethical considerations for assessment center operations. *Public Personnel Management*, **18**, 457–470.

Thornton, G.C. (1992) *Assessment Centers in Human Resource Management*. Reading, MA: Addison-Wesley.

Tziner, A., Ronen, S., Hacohen, D. (1993). A four-year validation study of an assessment center in a financial corporation. *Journal of Organizational Behaviour*, **14**, 225–237.

Zedeck, S. (1986) A process analysis of the assessment center method. *Research in Organizational Behaviour*, **8**, 259–296.