PROFESSIONAL FORUM:

A Process for Assessment Exercise Design: a Model of Best Practice

Yasmin Ahmed*, Tim Payne and Steve Whiddett

The use of assessment centres (ACs) has increased dramatically in the last five years. In a recent survey of 900 organizations, Wood, Boyle and Fullerton (1994) found that 45% of organizations with over 1,000 staff and 80% of organizations with over 4,000 staff used ACs. This is perhaps unsurprising given the amount of support for this method of assessment (e.g. Gaugler, Rosenthal, Thornton and Bentson 1987; Hunter and Hunter 1984; Sackett and Ryan 1985; Thornton and Byham 1982).

Whilst many organizations use 'off-the-shelf' assessment exercises, others prefer to develop bespoke exercises which are more suited to the particular organizational environment and culture, and more closely matched to the skills and abilities they wish to measure. The debate between the relative merits of the two types of exercises has tended to favour bespoke exercises (e.g. Adams 1987; Dulewicz 1991; Gratton 1985). Gratton argues that the use of generic, 'off-the-shelf' exercises contradicts the principles of AC technology. She suggests one of the main reasons for the widespread use of generic exercises is the lack of development costs associated with these exercises. We suggest another explanation for the use of generic exercises over bespoke exercises to be the absence of clear, step-by-step guidance which enables practitioners to construct exercises which conform to best practice. This adds a new dimension to Klimoski and Brickner's label of the AC as the 'modern enigma of human resource practice' (1987, p. 243).

Despite the great demand for bespoke AC exercises, there is surprisingly little guidance for designing exercises in either academic or practitioners' literature. Further, exercise design is not taught in many I/O Psychology Masters courses. A literature review revealed very few papers relevant to exercise design. No information was found specifically concerning beginning to end exercise design, however some information with design implications was found, mostly relating to the enhancement of the *content* and *construct* validity of exercises.

*Address for correspondence: Yasmin Ahmed, Tim Payne and Steve Whiddett, Pearn Kandola, 76 Banbury Road, Oxford OX2 6JT, UK.

Design for content

Several researchers, for example, Blanksby and Iles (1990); Goldstein, Zedeck and Schneider (1993); Iles (1992), Sackett (1987); Schneider and Schmitt (1992), Thornton and Byham (1982), and Whiddett and Branch (1993) outline rules of thumb for producing content valid exercises:

Ensuring exercises reflect the job and dimensions

- tasks should reflect the most important/ significant activities of the job
- the format used to present information in exercises should be the same as that experienced on the job (e.g. written/written; spoken/spoken)
- the way participants are asked to respond to exercises should match the way they would be expected to respond on the job (e.g. written vs spoken)
- the overall design of the centre is coherent, e.g. exercises are linked
- the tasks within each exercise are matched to the level of difficulty and complexity required in the job
- the form which the exercise takes should encourage appropriate behaviours, for example an interactive exercise to measure interpersonal sensitivity
- the exercise should reflect the organization's practices and culture.

Design for construct validity

In terms of maximizing the construct validity of exercises, most studies related to the exercise effect, and how to reduce it (Sackett and Dreher 1982). Several authors have made suggestions on how to reduce this effect (e.g. Gaugler and Thornton 1989; Iles 1992; Joyce, Thayer and Pond III 1994; Reilly, Henry and Smither 1990; Schneider and Schmitt 1992; Shore, McFarlane, Shore and Thornton 1992; Smith and Tarpey 1987; Whiddett and Branch 1993). All of the

suggestions focus on reducing the cognitive load on assessors by focusing on rating scale format and properties of dimensions.

Enhancing dimension definitions, scoring methods and rating scales

- dimensions should be carefully defined
- behaviours associated with dimensions should be observable in the exercises
- a small number of dimensions should be used overall
- there should be a small number in each exercise, ideally around three
- behaviourally anchored rating scales based on expert judgements should be used
- check-lists of the expected behaviours for each exercise should be provided for assessors
- original indicators from the job analysis should be used rather than exercise specific indicators
- all assessors should undergo thorough training.

This advice not withstanding, very little research evidence was found which contained specific guidance on how to design assessment exercises. This suggests that any design process adopted by experienced practitioners has until now remained implicit.

Research objective

The primary objective of the research was to identify and make explicit best practice in assessment exercise design.

Method

The sample

In–depth interviews were conducted with 14 practitioners. Of these, three were experienced occupational psychologists and the remainder were from a human resources background. Interviewees were currently employed in both the private and public sectors and were active in a range of business areas. All interviewees were experienced and active in the area of exercise design.

The interview method

In the first part of the interview, interviewees were asked to describe the process adopted in designing an exercise, from deciding how best to measure the dimensions through to evaluation of the exercise. Throughout the interview individuals were probed on their rationale for their approach and the resources and techniques which they used to aid design. The second part of the interview used critical incident technique (Flanagan 1954). Interviewees were asked to talk through design incidents and exercises which had gone well and those which had gone badly and the reasons for such an outcome. Typically, an interview lasted from two to two and a half hours and covered at least two types of exercise. The exercises discussed during interviews were role plays; in-basket; scheduling/planning exercises; business games; and group discussions (both assigned and non-assigned role). Two occupational psychologists content analysed the interview transcripts in terms of the process of design adopted for each of the different types of exercises.

Results

Despite the initial differentiation between exercise types, the analysis highlighted the commonality of the process used across all exercises. Thus, a generic model (Figure 1) emerged from the data comprising seven distinct stages.

The seven stages of design were identified from the interview data. Considerations at each stage were identified by both the interview data and the lessons drawn from the literature review. The model is not intended as a rigid, prescriptive

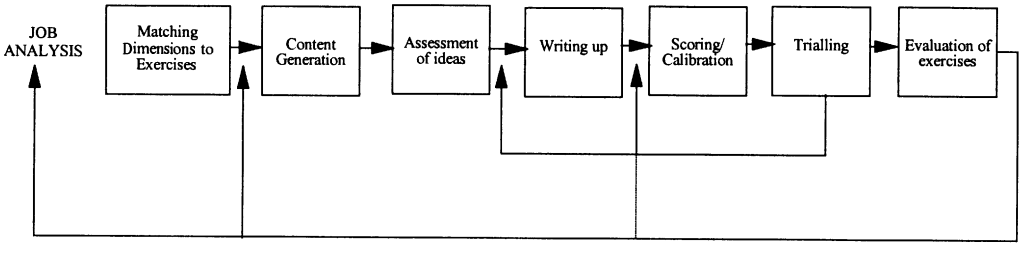

Figure 1: The practitioners' model of exercise design

Note: Job analysis is illustrated in the model since this ought to form the basis of any assessment method design. However, it is not represented as a stage in exercise design itself.

process but as a guide to structuring design. Indeed the model includes a number of feedback loops which highlight the dynamic nature of the process. Each stage is described in more detail below.

Stage 1 — Matching dimensions to exercises

During the first stage of design, interviewees described five main considerations. These were:

- *Form* — The form of the dimension greatly influenced the choice of an appropriate exercise. Designers chose quite different exercises to measure interactive dimensions, such as oral communication skills, compared to conceptual dimensions, such as analytical thinking.
- *Reflect job activities* — The choice of exercise was also governed by how closely it related to the job for which people were being assessed. As one interviewee stated, it makes little sense to ask supervisors to do a formal presentation if they do not do this in their actual job.
- *Acceptability* — The acceptability of the exercise to the participant was felt to be important. Practitioners tackled this aspect of design by ensuring the exercise clearly reflected the actual activities of the target role.
- *Number of dimensions per exercise* — Interviewees reported aiming to measure three dimensions and certainly no more than five dimensions per exercise in order to increase the likelihood of the accurate assessment of performance.
- *Multiple measures of dimensions* — Interviewees sought to measure each dimension at least twice in order to lessen the impact of a particularly poor performance on any one exercise affecting a participant's rating on a dimension. In practice, this was also found to be essential to ensure the full range of behaviours associated with a dimension were sampled and to reduce the potential of the exercise effect.

Stage 2 — Content generation

Interviewees reported that ideas for exercises generally came from the job analysis. However, when this information was not available, for example where the job did not yet exist, the two primary considerations in choosing a source of ideas were:

- *Representativeness* — The individual/s from whom designers seek ideas had to be representative of the target role in order to gain a realistic insight into the job. As a general rule, designers always spoke to several individuals to enhance the accuracy and comprehensiveness of the information.
- *Credibility* — A final criterion reported by some designers was whether or not the individuals are a credible source of information. This appeared to be a particular issue for exercises which were to be used on internal candidates.

Stage 3 — Assessment of ideas

Discussions with practitioners revealed seven common criteria for assessing the appropriateness of ideas. These were:

- *Importance to the job* — One of the key criteria for deciding the suitability of an idea was its relation to key outputs and responsibilities in the job.
- *Job vs. dimension related* — Interviewees reported some conflict between relating ideas to the job or the dimension. The result tended to be striking a balance between the dimension and the job. The rationale for this was to ensure that behaviours relating to the dimension were elicited whilst at the same time ensuring that the exercise reflected a likely scenario.
- *Practical constraints* — Some ideas were often rejected by designers simply because they were not translatable to exercises due to practical constraints e.g. time, cost, equipment.
- *Daily situation vs. unusual situations* — Generally designers used ideas which reflected reasonably typical scenarios associated with the target role. On occasions ideas which reflected unusual situations were used, but in moderation.
- *Levels of accomplishment* — Since one of the aims of assessment is to differentiate between different levels of performance, designers tended to choose ideas which could produce a task with different levels of accomplishment.
- *Fair* — A consideration which was reported to be of importance was the fairness of any idea. Designers described having to ensure that the chosen idea did not favour any particular individual due to their background, experience and so forth.
- *Politically acceptable* — A final consideration reported by some practitioners, was the need to ensure that the idea generated a politically neutral scenario.

Stage 4 — Writing up the exercise

Although this stage of design appears on the surface to be quite straightforward, interviewees described a variety of factors which influence the way the exercise is formatted.

- *Ensuring behaviour is elicited* — Designers reported the need to focus on the behaviour which s/he expects to result from the task. Some designers described trying to ensure eliciting the relevant behaviours by building dimension related aims/objectives into the exercise as part of the instructions.
- *Amount of material* — No quantitative rule appeared to exist in relation to the amount of material appropriate to exercises. Designers generally judged this in terms of what seemed manageable for the participant in the time available whilst at the same time reflecting reality and providing enough information to accomplish the task.
- *Clarity and comprehensiveness of information* — A consideration reported across all the interviews was the need to avoid any ambiguities which could detract from standardization and add to the anxieties of the participant.
- *Realism* — Ultimately assessment exercises are superficial. Designers reported the critical need to make the simulation as real as possible to ensure that the participant is engaged by the exercise.

Stage 5 — Scoring/calibration

This stage of design is critical to ensuring the construct validity of the exercise. Interviewees reported three main tactics used to ensure that participants' performance was accurately assessed.

- *Common rating scale vs. exercise specific rating scale* — In general, practitioners attempted to use just one type of rating scale throughout the assessment event. This certainly conforms to research findings which suggest that using just one scale reduces the cognitive load placed on assessors, enhancing construct validity.
- *Content related vs. dimension related* — Since the assessment of performance is in dimension terms, designers always reported relating guidelines directly to the dimensions. Where designers found that the focus was on content (inevitable in an exercise such as planning, and often in in-trays) the aim then became to make these directly translatable to the dimensions.
- *Distinguishing between the dimensions* — In line with the research findings regarding the exercise effect, designers were aware of the need to make the dimensions as distinguishable as possible. This was generally achieved by ensuring assessors were always provided with clear definitions of the dimension plus behavioural indicators, preferably specific to the exercise.

Stage 6 — Trialling the exercise

Interviewees described a range of issues which they examined when trialling exercises. These could be clustered into two categories:

- *Process issues* — These included; the timing of the exercise for the participants and the assessors; the clarity of the task; and how realistic/comfortable the exercise is for participants.
- *Content issues* — These included the level of difficulty; the clarity/amount of information; how realistic the exercise is to participants; how fair the exercise is to different participants; whether it elicits expected behaviours; whether it discriminates between different levels of ability.

Stage 7 — Evaluation of the exercise

In practice, interviewees were unable to report evaluation of the exercise in terms of a full validation study. The reason for this was one of practicality, in that the exercises they designed were often fairly specific and involved insufficient numbers of participants for a statistical validation. However, other types of evaluation were reported. These included:

- *Sitting in on wash-up sessions* — Designers were often keen to observe wash-up sessions to find out the type of evidence which was discussed and to highlight those behaviours which were not observed in the exercise.
- *Reviewing assessor sheets* — A similar reported technique was to examine the observer record sheets to ascertain the behaviour elicited by the exercise.
- *Assessor/participant reactions* — A useful source of ideas was found to be gauging the reactions of those who undergo the exercise, either as participant or assessor.

Discussion

The research represents the first study which has attempted to identify a systematic model of exercise design from start to finish. This seems a little ironic, since the expounded virtue of assessment exercises is their objectivity and standardization. However the dearth of information relating to design suggests that, to a degree, exercise design has been conducted in a somewhat *ad hoc* manner. This research has identified a number of underlying guiding principles practised by experts when designing assessment exercises. These themes can be grouped into a coherent model which is influenced by design practice and the assessment literature. We believe that the model represents a

key contribution to the practice of assessment exercise design in four ways.

First, the model provides guidance for designing assessment exercises. We believe there to be a clear need for such guidance given the lack of practical advice offered by the assessment literature and also academic courses. The model identified in this study capitalizes on the expert knowledge of designers of assessment exercises whilst also incorporating findings from the literature. This is likely to be valuable both to novice designers and experienced designers in providing a range of considerations at each stage of exercise development to ensure best practice is met.

Second, by making criteria for best practice explicit at every stage of design, a framework is created to continuously evaluate exercises. Evaluation has traditionally occurred following design. Although evaluation is represented as the seventh stage of design, we would advocate that evaluation of the exercise should be an inherent aspect of the entire process of exercise design. The model facilitates this process by highlighting the many considerations, which if overlooked, could compromise the quality of the exercise. It offers an iterative process where considerations at particular stages may lead the designer to re-visit earlier stages and design quality into the exercise.

Third, the model can be used to evaluate retrospectively the design of assessment procedures. Traditional evaluations of assessment centres and exercises has tended to focus on predictive validity studies which have sought to establish whether or not a statistical relationship exists between exercise scores and on the job performance. This type of study is valuable to ascertain whether or not the procedure predicts success in the job, however the information that it yields is very narrow. Further, a relatively large sample of job applicants is required to conduct this type of study which limits its application. However, reliance on this method of evaluation can be misguided. There are many variables in a selection procedure which impact on the overall relevance and validity of the procedure. Validity is not a characteristic of a test or assessment exercise, it is an aspect of its use. It is essential to scrutinize the way exercises are intended to be used, the way and the contexts in which they are used, as well as how and for what purpose they were designed.

The framework provided by the model can be used to evaluate procedures. The authors have devised a number of 'tools', (for example, checklists, matrices, rules of thumb and blueprints) which are derived from the model and enable an evaluation of the fairness, reliability and construct and content validity of exercises as well as guiding the designer when developing exercises (Whiddett, Boyle, Payne and Ahmed 1996). This type of evaluation provides a method which can be applied to a wider variety of situations, for example where only a small sample of data is available. Further, the method provides detailed information which can highlight where problems exist in the process and so, where to focus attention to rectify.

Fourth, the use of the model and its associated techniques not only provide a clear design route for practitioners, but can also save valuable time by 'streamlining' the whole process without compromising best practice. We tested this out during an exercise design workshop on a group of twelve Human Resources professionals, varying in experience of exercise design, from complete novices to experience of several years. Over the period of a morning, all the participants were able to come up with a design for a role play exercise, instructions and a scoring mechanism. This goes some way to countering the argument of bespoke exercises being costly in time.

The model represents a systematic process of assessment exercise design. We hope that the model will make a valuable contribution to exercise development by providing a clear structure which will give rise to standardized exercises developed through a standardized process which conforms to best practice.

References

Adams, D. (1987) Assessment centre exercises—Bespoke or ready to wear? *Guidance and Assessment Review* **3**(1).

Blanksby, M and Iles, P. (1990) Recent developments in assessment centre theory, practice and operation. *Personnel Review* **19**(6) 33–40.

Dulewicz, V. (1991) Improving assessment centres. *Personnel Management* June.

Flanagan, J.C. (1954) The critical incident technique. *Psychological Bulletin* **51**, 327–58

Gaugler, B.B., Rosenthal, D.B., Thornton G.C. and Bentson, C. (1987) Meta-analysis of assessment center. *Journal of Applied Psychology*, **72**(4) 611–618.

Gaugler, B.B. and Thornton III, G.C. (1989) Number of assessment centre dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology*, **74** (4) 611–618.

Goldstein, I., Zedeck, S. and Schneider, B. (1993) An exploration of the job analysis-content validity process. In N. Schmitt, W.C. Borman Associates (eds.) *Personnel selection in organisations*. San Fransisco: Jossey Bass.

Gratton, L. (1985) Assessment Centres: Theory, research and practice. *Human Resource Management Australia*, **23**, 10–14.

Hunter, L.E. and Hunter, R.F. (1984) Validity and

utility of alternative predictors of job performance. *Psychological Bulletin*, **96**, 72–98.
Iles, P. (1992) Centres of excellence? Assessment and development centres, managerial competence and human resources strategies. *British Journal of Management*, **3**, 79–90.
Joyce, L.W., Thayer, P.W. and Pond III, S.B. (1994) Managerial functions: an alternative to traditional assessment centre dimensions? *Personnel Psychology*, **47**(1) 109–122.
Klimoski, R. and Brickner, M. (1987) Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, **40**, 243–260.
Reilly, R.R., Henry, S. and Smither, J.W. (1990) An examination of the effects of using behavioural checklists on the construct validity of assessment centre dimensions. *Personnel Psychology*, **43**, 72–84.
Sackett, P.R. (1987) Assessment centres and content validity: Some neglected issues. *Personnel Psychology*, **40**, 13–25.
Sackett, P.R. and Dreher, G.F. (1982) Constructs and assessment centre dimensions: some troubling empirical findings. *Journal of Applied Psychology*, **67**, 401–410.
Sackett, P.R. and Ryan, A.M. (1985) A review of recent assessment centre research. *Journal of Management Development*, **4**(4) 13–27
Schneider, J.R. and Schmitt, N. (1992) An exercise design approach to understanding assessment centre dimensions and exercise constructs. *Journal of Applied Psychology*, **77**(1) 32–41.
Shore, T.H., McFarlane Shore, L. and Thornton III, G.C. (1992) Construct validity of self- and peer evaluations of performance dimensions in an assessment centre. *Journal of Applied Psychology*, 72(1) 42–54.
Smith, D. and Tarpey, T. (1987) In-tray exercises and assessment centres: The issue of reliability. *Personnel Review*, 16(3) 24–28.
Thornton, G.C. III and Byham, W.C. (1982) *Assessment Centres and Managerial Performance*. London: Academic Press.
Whiddett, S, and Branch, J. (1993) Development centres in Volvo. *Training and development*, **November**, 16–18
Whiddett, S., Boyle, S., Payne, T. and Ahmed, Y. (1996) *Tools for Assessment and Development Centres* London: Institute of Personnel and Development.
Wood, R., Boyle, S., and Fullerton, J. (1994) The competencies that organisations are looking for in assessment centres. *Competency*, **2**(1) 32–34.

Bibliography

Anderson, N.R., Payne, T., Ferguson, E. and Smith, T. (1994) Assessor decision making, information processing and assessor decision strategies in a British assessment centre. *Personnel Review*, **23**(1) 52–62.
Bedford, T. (1987) New Developments in Assessment centre design. *Guidance and Assessment*, **June, 3**(3).
Boyle, S., Fullerton, J. and Yapp, M. (1993) The rise of the assessment centre: A survey of AC usage in the UK. *Selection and Development Review*, **9**(3) 1–4.
Feltham, R. (1992) Assessment Centre decision making: judgemental vs mechanical. *Journal of Occupational Psychology*, **61**, 237–241.
Gaugler, B.B. and Rudolph, A.S. (1992) The Influence of assessee performance variation on assessors' judgements. *Personnel Psychology*, **45**, 77–90.
Hakstian, A.R. and Harlos, K.P. (1993) Assessment of in-basket performance by quickly scored methods: Development and psychometric evaluation. *International Journal of Selection and Assessment*, **1**(3) 135–142.
Hakstian, A.R., Woolley, R.M. and Woolsey, L.K. (1991) Management selection by multiple domain assessment: concurrent validity. *Educational and Psychological Measurement*, **51**, 883–898
Hakstian, A.R., Woolsey, L.K. and Schroeder, M.L. (1986) Development and application of a quickly scored in-basket exercise in an organisational setting. *Educational and Psychological Measurement*, **46**, 385–396.
Highhouse, S. and Harris, M.M. (1993) The measurement of assessment centre situations: Bem's template matching technique for examining exercise similarity. *Journal of Applied Social Psychology*, **23**(2) 140–155.
Harris, M.M., Becker, A.S. and Smith, D.E. (1993) Does the assessment center scoring method affect the cross-situational consistency of ratings? *Journal of Applied Psychology*, **78**(4) 675–678.
Jansen, P. and Stoop, B. (1994) Assessment centre graduate selection: decision processes, validity and evaluation by candidates. *International Journal of Selection and Assessment*, **4**(2) 193–208.
Jones, A., Herriot, P., Long, B. and Drakely, R. (1991) Attempting to improve the validity of a well established assessment centre. *Journal of Occupational Psychology*, **64**, 1–21.
Kesselman, G.A., Lopez, F.M. and Lopez, F.E. (1982) The development and validation of a self-report scored in-basket test in an assessment centre setting. *Public Personnel Management Journal*, **11**, 228–238.
Kleinman, M. (1993) Are rating dimensions in assessment centres transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology*, **78**(6) 998–993.
Lowry, P.E. (1992) The assessment center: Effects of varying consensus procedures. *Public Personnel Management*, **21**(2) 171–183.
Payne, T., Anderson, N.R. and Smith, T. (1992) Assessment centres, selection systems and cost effectiveness: An evaluative case study. *Personnel Review*, **21**(4) 48–56.
Pynes, J. and Bernardin, H.J. (1992) Mechanical vs consensus derived assessment center ratings: a comparison of job performance validities. *Public Personnel Management*, **21**(1) 17–28.
Robertson, I., Gratton, L. and Sharpley, D. (1987) The psychometric properties and design of managerial assessment centres: dimensions into exercises won't go. *Journal of Occupational Psychology*, **60**, 187–195.
Silverman, W.H., Dalessio, A., Woods, S.B., Johnson, R.L. (1980) Influence of assessment centre methods on assessors' ratings. *Personnel Psychology*, 39, 565–578.

Stamoulis, D.T., and Hauenstein, N.M.A. (1993) Rater training and rater accuracy: training for dimensional accuracy versus training for rate differentiation. *Journal of Applied Psychology*, **78**(6) 994–1003.

Tenopyr, M.L. (1977) Content-construct confusion. *Personnel Psychology*, **30**, 47–54.

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

Watson, W.E. and Behnke, R.R. (1990) Group identification, independence and self-monitoring characteristics as predictors of leaderless group discussion performance. *Journal of Applied Social Psychology*, **20**(17) 1423–1431.