Journal of Occupational and Organizational Psychology (1995), 68, 25–47 *Printed in Great Britain*

Two field tests of an explanation of assessment centre validity

Craig J. Russell*

Department of Management, College of Business Administration, Louisiana State University, Baton Rouge, LA 70803-6312, USA

Donald R. Domm

Department of Management, John Carroll University

Klimoski & Brickner (1987) described two sets of constructs underlying assessment centre ratings. The trait explanation holds that dimensional ratings capture a candidate's personal characteristics, skills and abilities. The performance consistency/role congruency explanation holds that dimensional ratings are predictions of how well the candidate will perform various tasks and/or roles in the target job. While past research has failed to find support for the trait explanation, no studies have explicitly examined the validity of assessment centres designed to make task or role-based dimensional ratings. We report two field evaluations of this explanation. In Study 1 assessor training was modified to have assessors view traditional assessment dimensions as role requirements. Concurrent validation of assessor evaluations of retail store managers resulted in correlations ranging from .22 to .28 with superiors' performance appraisal ratings and .32 to .35 with store profit. Study 2 evaluated the criterion-related validity of ratings on both job requirements and traits. Findings indicate that task-based ratings demonstrate concurrent validity in a sample of entry level unit managers while the traditional trait-based ratings do not. Implications for the construct validity and design of assessment centres are drawn.

Klimoski & Brickner (1987) described six alternative explanations for the construct validity of assessment centre ratings. The traditional trait explanation (Byham, 1970, 1980; Holmes, 1977; Standards for Ethical Considerations for Assessment Center Operations, 1977), that assessment centre dimensional ratings capture individual differences in candidates' skills and abilities, has been the subject of numerous empirical investigations. For example, when assessment centres are designed to yield trait ratings of dimensions immediately after each exercise, ratings of the same dimension are expected to be highly correlated with one another regardless of the exercise. Post-exercise dimensional ratings would also be expected to yield low correlations between ratings of different dimensions obtained within the same exercise. Bycio, Alveres & Hahn (1987); Neidig, Martin & Yates (1979); Russell (1987); Sackett & Dreher (1982); Shore, Thornton & Shore (1990);

*Requests for reprints.

and Turnage & Muchinsky (1982) reported mixed evidence of convergent validity within dimensions across exercises and no evidence of discriminant validity within exercises. Perhaps most discouraging, Russell (1987) and Shore *et al.* (1990) failed to find convergent validity between post-exercise dimensional ratings and independent measures of candidate traits obtained outside of the assessment centre. Evidence to date strongly suggests that post-exercise dimensional ratings are not construct valid representations of candidates' personal characteristics, skills and abilities. This does not mean assessment ratings are not construct valid—there must be some systematic variance captured in the ratings otherwise criterion-related validity would not occur. Simply put, assessment centre ratings must be valid representations of some construct(s), we just do not know which one(s).

Unfortunately, evidence shedding light on constructs underlying ratings made in other assessment centre configurations is nonexistent. There is no evidence suggesting that assessors following the American Telephone and Telegraph prototype (where evaluation of dimensions is postponed until all exercises have been completed) are evaluating the same constructs as assessors making post-exercise evaluations. Further, no evidence has been reported bearing on any of the five alternatives Klimoski & Brickner (1987) presented to the 'traditional' view of assessment centre construct validity. Perhaps the simplest alternative is what they called the 'performance consistency' explanation, where assessors are evaluating candidates' task performance 'thus bypassing the judgment of traits entirely' (Klimoski & Brickner, 1987, p. 249). Russell (1987) labelled this the role congruency explanation, hypothesizing that assessors evaluate candidate behaviours in terms of role or performance requirements in the target position. The performance consistency/role congruency explanation implies that these constructs reflect performance on latent task characteristics of the target job. The purpose of the two studies reported in this article is to test hypotheses regarding the construct and criterion-related validity of assessment centres designed to measure job performance constructs. We will briefly discuss the nature of these constructs before describing the hypotheses to be examined.

Numerous job-oriented constructs are consistent with the performance consistency/role congruency explanation. Assessors may view each assessment exercise as a meaningful component of the job performance construct domain, arriving at post-exercise ratings that reflect the quantity and quality of candidate behaviours exhibited in each exercise. Alternatively, given their importance in assessor training, assessors may view each dimension as the most meaningful component of the job performance construct domain. For example, assessors may view behaviours associated with the dimension 'initiative' as meaningful predictors of performance in the 'initiative' component of the target job (and not the amount of 'initiative' possessed by the candidate).

The nature of any job-specific construct may depend on the assessment centre configuration in use. For example, when dimensional ratings are made after an in-basket exercise, the unique profile of tasks imbedded in that exercise might cause all dimensional ratings based on that exercise to be highly correlated with one another (an exercise halo effect) and less correlated with dimensional ratings made after an interview simulation exercise. Conversely, when dimensional ratings are made after observing performance on all exercises (the AT&T prototype), the tasks imbedded in any specific exercise may be less salient. Using the earlier example, assessors in this prototype may view dimension definitions as a job-based organizing heuristic where definition of the 'initiative' dimen-

sion is conceived in terms of the initiative requirements in the entire target job domain (not just tasks embedded in a single exercise). Relevant behaviour observations across all exercises are then considered in arriving at an 'initiative' rating. Clearly, research is needed to examine how variations in the assessment process (exercise content, exercise construction, rating procedures, assessor training, etc.) impact assessors' ability to make valid ratings of person- or job-oriented constructs.

Our purpose is to present results from two studies in field settings that examine a performance consistency/role congruency explanation of dimensional ratings made in a centre modelled after the AT&T procedures. To our knowledge, these are the first studies of assessment centre construct validity in which the *a priori* centre design has not been based on traditional conceptualization, i.e. assessors are not trained to view dimensional ratings as measures of candidates' personal characteristics, skills and abilities.

Study 1 made only one modification of the traditional AT&T prototype of assessment design, in which assessors were explicitly trained to view traditional assessment centre dimensions as role requirements of the target position. All other aspects of the assessment design conformed with traditional blueprints of centres designed for selection purposes (Standards for Ethical Considerations for Assessment Center Operations, 1977). Evidence of criterion-related validity could be considered preliminary support for the hypothesis that assessors can use constructs found in the job performance domain to arrive at dimensional ratings.

Study 2 considered an additional modification where assessors were trained to obtain both traditional dimensional ratings of candidate skills, abilities, and personal characteristics and forecasts of how well candidates should perform on various tasks in a future job. This permits a direct test of the competing explanations of assessment centre construct validity in an assessment centre modelled after the AT&T design—task-based dimensional ratings are hypothesized to demonstrate higher criterion-related validities if assessors are making valid evaluations of candidate performance on job-based constructs, while the reverse should occur if assessors are making construct valid evaluations of candidates' traits. Given prior research efforts which have failed to find support for a trait-based explanation, we hypothesize the task-based ratings will demonstrate higher criterion-related validities than the trait-based ratings. All other aspects of Study 2 assessment centre design were consistent with the Standards for Ethical Considerations for Assessment Center Operations (1977).

STUDY 1

Method

Subjects

The host firm, a *Fortune* 100 vertically integrated durable goods manufacturer, wanted an assessment centre to select store managers for the approximately 900 retail outlets operated throughout the United States. One randomly selected set of 140 current store managers attended a one-day assessment centre. Twenty-five top performing district managers (who did not know any of the assessees) were trained as assessors—all had been store managers earlier in their careers. Another sample consisted of direct superiors (district managers) of all store managers evaluated in the assessment centre. These district managers provided performance appraisal ratings on each store manager evaluated in the assessment centre.

Job analysis

A job analysis inventory was designed to insure that the job domain of task and behavioural requirements had been adequately sampled for development of assessment centre exercises and a performance appraisal instrument. This is a necessary but not sufficient step for construct valid ratings of role requirements in the target job (cf. Tenopyr, 1977). Structured interviews were conducted with 70 current store managers in order to identify store manager goals and objectives, the tasks that had to be completed in order to achieve each goal and objective, and the specific behavioural duties that had to be completed in order to yield successful task accomplishment.

The information obtained from the structured interviews was distributed to corporate human resource management personnel, select district managers (store managers' direct superiors), and higher level management for modifications, additions, and deletions. These subject matter experts initially grouped 17 tasks emerging from the structured interviews into the type of organizational resource being managed [labelled personnel responsibilities (labour resources), resource management responsibilities (raw material resources), customer interaction and corporate citizenship (customer and corporate resources), and operations management responsibilities (information resources) in Table 1]. They also noted that operations management responsibilities tended to be dominant and in fact permeate the other performance dimensions (i.e. managing information about raw materials, customers, etc., makes it implicitly related to the other performance dimensions). Copies of the job analysis inventory are available from the first author.

Information from the structured interviews was used to develop a job analysis inventory targeting the importance and frequency of job tasks. Inventory forms were completed by 450 current store managers and

Table 1. Resource responsibilities and corresponding tasks

I. Personnel responsibilities—obtaining, maintaining and motivating store employees
   1. Recruiting/selection
   2. Training
   3. Setting objectives and communicating responsibilities
   4. Performance appraisal
   5. Ongoing communication and motivation
   6. Career guidance

II. Resource management responsibilities—the planning, organization and control of inventory, tools, equipment and facility
   7. Inventory control
   8. Equipment maintenance
   9. Cash control
   10. Cost control

III. Customer interactions and good corporate citizenship—activities at a local level that generates a desirable public image so as to enhance the store's sales and profit objectives
   11. Customer relations
   12. Selling
   13. Maintaining company image

IV. Operations management responsibilities—record keeping/paperwork, audit rating, merchandising, advertising, communicating with assistant district manager
   14. Record keeping and paper work
   15. Merchandising
   16. Advertising
   17. Communication with district manager

returned directly to the investigators. Item analyses indicated that almost all items received an average importance rating greater than 2.5 (the scale mid-point) and an average frequency rating greater than 3.0 (the scale mid-point). Those few items that had mean ratings lower than the scale mid-point for importance had ratings higher than the scale mid-point on the frequency rating and vice versa. Factor analyses were not performed due to Cranny & Doherty's (1988) demonstration that it is impossible to interpret factor structures of importance and frequency ratings. Hence, the subject matter experts' resource management categories were preliminarily viewed as meaningful groupings of tasks deemed important for successful performance as a store manager. All aspects of the resource management categories were woven into the assessment centre exercises.

Development of assessment centre exercises

Construction of the current exercises was performed by the authors based on the outcome of job analysis procedures. Preliminary drafts of exercise materials were distributed for review to district managers, selected line managers at the corporate headquarters, and corporate human resource personnel to ensure that all aspects of the job were appropriately sampled. Modifications in exercise materials resulted in two iterations of this process before yielding the final exercises.

While tasks from all resource management categories were woven into each exercise, exercises tended to be dominated by one or more particular type of responsibility. Specifically, the customer interaction job requirements were used to construct an interview simulation exercise; personnel responsibility job requirements contributed to the construction of the in-basket and leaderless group exercises; while corporate citizenship, personnel, resource management and operations management responsibilities contributed to the in-basket exercise.

Interview simulation

The interview simulation was designed to give candidates the opportunity to deal with a disgruntled customer (Crooks, 1977). The candidate was provided with partial information about a customer who had made an appointment to talk about a problem. The customer was a trained role player who had more information than the candidate. The candidate's job was to gather information from the customer, develop possible alternative plans to deal with the customer's problem, and generally keep the customer happy. The role player was trained to offer minimal information (unless explicitly asked) and remain unhappy regardless of what the candidate said or did.

Leaderless group exercise

Bass (1949) demonstrated that behaviour can be systematically observed and recorded to capture relevant performance dimensions in leaderless group exercises. Candidates were assessed in groups of six. Within a cohort, each was asked to play the role of a store manager. Each candidate was given a different profile of one of his/her subordinates. The candidates' task was to identify the best two subordinates for two new district commercial sales positions. Candidates had to deal with a mild role conflict—they had to identify the best two subordinates to promote into commercial sales for the district while simultaneously trying to get their own subordinate promoted to ensure that the majority of any new sales accrued to their own store.

In-basket exercise

The in-basket exercise contained 18 pieces of information simulating problems a newly appointed store manager might encounter on the job (Hemphill, Griffiths & Frederiksen, 1962). Candidates were interviewed about how and why they dealt with each item after working on the exercise for 1.5 hours. The items were constructed to reflect tasks in the categories of personnel, resource management and operations management responsibilities.

Behaviours observed in these assessment centre exercises could be used to draw inferences about characteristics of the individual *and/or* make forecasts about performance in different roles required by a particular job. After development of the exercises, eleven dimensions (see Table 2) were chosen that were representative of

Table 2. Performance consistency/role congruency assessment centre dimensions in Study 1

Energy: Degree to which behaviours meet the continuously high level of work activity required of the job.
Forcefulness: Degree to which behaviours commanded the attention and had an impact on others as required on the job.
Initiative: Degree to which behaviours influence events to achieve goals by originating action rather than merely responding to events as required on the job.
Impact: Degree to which presentation of self makes a favourable impact on others as required on the job.
Organization and planning: Degree to which work is effectively organized and planned for the future as required on the job.
Decisiveness: Degree to which decisions are made as required on the job.
Judgement: Degree to which decisions of high quality are made as required on the job.
Social sensitivity: Degree to which subtle cues are perceived in the environment as required on the job.
Behaviour flexibility: Degree to which behaviour is modified to reach goals as required on the job.
Leadership: Degree to which appropriate interpersonal styles and methods are used in guiding individuals toward task accomplishment as required on the job.
Oral communication: Degree to which thoughts and ideas are conveyed in a clear, unambiguous, and effective manner as required on the job.

those used in traditional assessment centre procedures (Bass, 1949; Crooks, 1977; Hemphill *et al.*, 1962; and Howard & Bray, 1988). However, unlike previous applications which focused definitions on person characteristics, skill and abilities, the current dimension definitions focused on the use of behavioural observations as representations of performance and role requirements in the target job. This focus was primarily instilled through assessor training (though the dimension definitions contain a distinct job-oriented focus).

Assessor training

Assessor training took place in one day. After a brief description of what an assessment centre was, assessors were given a short lecture on rating error and the definition of assessment dimensions. All dimension definitions described groups of behaviours *required of the job.* Descriptions of the training activities can be obtained from the first author. The major emphasis during assessor training was on the distinction between tasks, behaviours and other factors like personality characteristics or traits that an assessor might feel have a bearing on assessment centre performance. Assessors were informed that dimension labels like 'energy' or 'social sensitivity' could refer to either characteristics of the person or requirements of the job and that the latter would be emphasized in the assessment centre. The importance of observing and identifying behaviours required of the job and not making inferences about what kind of person the candidate might be was repeated in the context of each exercise and the consensus discussion. Repeated comparisons were made to the performance appraisal distinction between evaluations used to identify developmental needs (i.e. evaluations of skills and abilities) vs. evaluations used to make actual personnel decisions (e.g. promotion, merit pay) based on task performance. Questions asked by assessors over the course of the day indicated they understood the distinction between rating the person vs. rating how well the person will do what is expected of him or her.

For the remainder of the training day assessors went through role plays, where some assessors were candidates while others observed and rated performance, and instruction on rating error in the context of assessment centre observations. Again, special attention was focused on the fact that ratings *did not* represent personal traits, such as being a socially sensitive person. Instead, ratings were to indicate how well the candidate's behaviour met the social sensitivity requirements of the job. Many of the assessors indicated this distinction was intuitively appealing in light of concerns they had about detecting faking. Specifically, the question was raised in the various role plays about how to know if a candidate was faking to get a high score. Assessors were told that, assuming all candidates are trying to do their best, an assessment centre measures a candidate's maximum performance and that some candidates may try to engage in a non-typical array of behaviours. However, assessors were also told that candidates cannot fake behaviour(s) that they do not have the capacity to perform. Assessors were assured that any slippage due to, for example, unmotivated and lazy underachievers putting on a good face for assessment would be captured elsewhere in the selection process (e.g. promotion recommendations by superiors). Over the course of the role plays many assessors expressed relief that they did not have to evaluate the 'person' as they felt they could not justify any insights into the true nature of candidate.

Rating process

Three assessors observed and evaluated each assessee. After gathering all observations on the three exercises and individually arriving at ratings on all the assessment dimensions, assessors presented their ratings to one another in a group discussion. In case of disagreement, discussion took place until consensus was reached. Discussion was based entirely on the observations made and how the observations were evaluated. Discussion took place until the assessors were within at least one rating point of each other, at which time a vote was taken and recorded as consensus. Russell (1985) and Sackett & Hakel (1979) have demonstrated that the vast majority of variance in overall assessment centre ratings could be explained by a simple sum of the dimension ratings. However, it was felt that assessors' clinical combinations of the assessment centre dimensions might capture something a simple sum of the dimensions would not. Consequently, the last step assessors took was to combine the individual assessment dimension ratings into an overall assessment rating.

Performance criteria

Performance appraisal information was gathered on task-level performance ratings and store-level financial performance. Supervisors (district managers) were asked to rate each of their store managers on the 17 tasks identified within each of the four categories of responsibilities. Behavioural requirements taken from the job analysis were listed below each task. At the end of each of the resource responsibility categories, district managers were asked to provide an overall performance rating for the category. None of the district managers (indeed, none of the firms' personnel other than the assessors) were aware of the assessment centre ratings. All performance appraisal forms were returned directly to the investigators. A copy of the performance appraisal form is available from the first author.

Finally, data on gross sales, net profit, and store size (number of employees) were made available for 87 of the 140 stores. Financial performance data on each store were not centrally located within the firm (data were kept only at the district level) and limited investigator resources prevented anything more than a two-week search of decentralized archival records.

Results

Reliability on performance appraisal data

Internal consistency reliability coefficients (Cronbach's coefficient alpha) were calculated on task ratings made within each resource responsibility category. The coefficients were as follows: personnel responsibilities (6 task ratings, $\alpha = .88$); resource management responsibilities (4 task ratings, $\alpha = .90$); customer interactions and corporate citizenship (3 task ratings, $\alpha = .79$); and operations management responsibilities (4 task ratings,

$\alpha = .85$). All internal consistency reliability coefficients indicate more than adequate amounts of systematic (non-random) variation in the performance ratings.

Psychometric properties of the assessment centre and performance appraisal

A major psychometric property of interest in the current assessment centre ratings is the degree to which the overall assessment rating is reliably reached by assessors. To examine this question, the overall ratings were regressed onto the 11 assessment centre dimensional ratings ($N = 140$). An R^2 or .84 means that assessors were basing the majority of their overall rating decision on the observations that went into the assessment ratings (all dimension ratings generated significant regression coefficients). Common factor analysis with varimax rotation performed on dimensional ratings indicated that one factor was dominant (first factor had an eigenvalue of 6.2 and an average loading of .75, while the second factor had an eigenvalue of 1 and average loading of .27). These results are congruent with findings reported by Russell (1985) and Sackett & Hakel (1979).

Common factor analysis with varimax rotation was performed to examine how well the performance appraisal ratings conformed with the *a priori* four resource categories. Only three factors had eigenvalues greater than one. Clean, interpretable loadings emerged when the performance appraisal ratings were forced onto three factors (loading rule of at least .50 on the major factor and .40 on all other factors—average loading on major and minor factors were .73 and .17, respectively). Factor 1 contained all ratings made on human resource responsibilities, factor 2 contained all customer relations/corporate citizenship responsibility ratings, and factor 3 contained all resource management responsibility ratings. Operations management responsibilities had its four task ratings load on each of three factors. Paperwork/record keeping loaded at .50 on factors 1 and 3 (apparently paperwork is most highly associated with human resource and materials management) while merchandising loaded on factor 2 (customer relations/corporate citizenship responsibilities). It appeared that performance of operations management responsibilities was an integral part of store managers' performance of other aspects of their jobs, confirming information provided by the subject matter experts. Thus, factor loadings confirmed at least three distinct performance dimensions identified from the job analysis.

Criterion validity analyses

All correlations among the predictors (11 assessment centre dimensional ratings and the overall assessment centre rating) and the performance criteria are reported in Table 3. For purposes of brevity, means, standard deviations, and correlations are only reported for the overall ratings made for each performance appraisal resource category. In addition, an overall performance rating (PASUM) was calculated by summing the overall personnel, resource management, customer relations/corporate citizenship and operations management responsibility ratings. Moderate correlations were found between assessment centre dimensional ratings and performance ratings. The overall assessment rating correlated with the sum of overall responsibility ratings (PASUM) at $r = .28$ ($p \leq .01$) and .32 ($p \leq .01$) with store profit. Candidates who received an overall assessment rating of 3 or 4 (on the 4-point OAR scale) generated an average of over $3 000 more quarterly profit at their stores relative to candidates who received ratings of 1 or 2. To our knowledge, this

Table 3. Means, standard deviations and correlations among predictors and criterion ($N = 140$ for all rows except 19 and 20, where $N = 87$)

Variables	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Energy	3.63	.88	–																		
2. Forcefulness	3.33	.93	.71**	–																	
3. Initiative	3.5	.83	.59**	.74**	–																
4. Impact	3.35	.87	.56**	.65**	.54**	–															
5. Organization and planning	3.32	1.03	.44**	.50**	.58**	.56**	–														
6. Decisiveness	3.57	.75	.43**	.47**	.54**	.48**	.61**	–													
7. Judgement	3.20	.93	.53**	.60**	.53**	.54**	.55**	.54**	–												
8. Social sensitivity	3.39	.94	.43**	.37**	.46**	.49**	.47**	.58**	.53**	–											
9. Behaviour flexibility	3.22	.85	.42**	.50**	.52**	.43**	.42**	.52**	.56**	.65**	–										
10. Leadership	3.41	.87	.69**	.75**	.65**	.65**	.46**	.46**	.66**	.51**	.53										
11. Oral communication	3.75	.81	.67**	.62**	.54**	.50**	.41**	.36**	.57**	.48**	.48**	.67**	–								
12. Overall assessment rating	3.15	.85	.71**	.70**	.65**	.65**	.70**	.60**	.64**	.57**	.58**	.75**	.64**	–							
13. ACSUM[a]	37.76	7.41	.78**	.82**	.80**	.77**	.74**	.71**	.79**	.71**	.71**	.86**	.74**	.86**	–						
14. Personnel responsibility	3.23	.84	.21*	.13	.12	.12	–.02	.00	–.01	.00	.11	.14	.22	.16	.12	–					
15. Resource management responsibility	3.56	.78	.02	.08	.05	.08	.13	.05	.09	–.07	.04	.10	.13	.20*	.08	.38**	–				
16. Customer relations responsibility	3.92	.81	.18*	.13	.06	.24**	.21*	–.04	.16*	.17*	.27**	.23**	.27**	.27**	.23**	.27**	.24**	–			
17. Operations management responsibility	3.57	.70	.11	.11	.11	.18*	.17*	.02	.01	.05	.08	.12	.07	.21*	.12	.56**	.45**	.45**	–		
18. PASUM[a]	14.23	2.25	.12	.11	.08	.20*	.17*	.00	.08	.05	.17*	.20*	.22*	.28**	.16	.76**	.69**	.67**	.80**	–	
19. Sales[a]	$285736.59	115893.42	.00	–.05	.06	.10	–.01	–.06	–.10	–.02	.12	–.07	–.07	.05	–.02	.24*	–.02	.35**	.35**	.33*	–
20. Profit[a]	–$3662.69	16902.05	.18	.32**	.26**	.34**	.33**	.29**	.32**	.12	.26**	.34**	.23*	.32**	.35**	.16	.20*	.14	.18	.24*	.33**

* $p \leq .05$, two-tailed test;
** $p \leq .01$, two-tailed test.
a: ACSUM = the simple sum of all assessment centre dimensional ratings; PASUM = simple sum of all overall responsibility ratings; Sales = gross sale over the first four months of 1989, though correlations were derived after controlling for store size (# of employees); Profit = net profit over first four months of 1989, though correlations were derived after controlling for store size.

is the only reported evidence of a relationship between assessment centre ratings and a firm's financial performance. The average correlation between the 11 dimensional ratings and the performance ratings within each responsibility area (not reported in Table 3) was .30, $N = 44$, SD = .05. The entire correlation matrix is available from the authors on request.

Multiple regression analyses were performed to examine how well combinations of assessment centre ratings predicted performance. The average multiple correlation between the 17 individual performance appraisal ratings and assessment centre ratings was .36. The multiple correlation between PASUM and the assessment centre ratings was .40. This is comparable with the average validity coefficient reported in the most recent meta-analysis of assessment centre studies of .38 (Gaugler, Rosenthal, Thornton & Bentson, 1987), though the multiple correlations reported here are probably slightly inflated due to random chance and would shrink if subjected to cross validation. However, unlike multiple regression techniques, a simple sum of the 11 dimensional ratings (ACSUM) would not take advantage of sampling error. ACSUM was significantly correlated with customer relations and corporate citizenship responsibilities ($r = .23, p < .01$) and profit ($r = .35, p < .01$) when profit was corrected for store size. Regression of profit onto the 11 assessment centre dimensional ratings resulted in $R = .48$ ($p < .05$). Assessment centre ratings were not significantly related to sales volume after adjusting it for store size.

Discussion

These results were derived from an AT&T prototypic assessment centre that had been modified in only one way. In contrast to traditionally designed assessment centres, the current procedure adopted the theoretical position that dimensional ratings reflect how consistent candidates' behaviours were with performance and/or role requirements in the target position. The performance consistency/role congruency procedure avoids Sackett's concerns about the complexity of assessor rating processes (Sackett, 1982; Sackett & Harris, 1988). These ratings yielded uncorrected criterion validities that compared favourably with criterion validities corrected for sampling and measurement error reported in traditional assessment centres (Gaugler *et al.*, 1987). The results are the first to suggest that a performance consistency/role congruency explanation can account for assessment centre criterion-related validities.

Interestingly, examination of descriptive statistics indicated the average profit of the stores in the sample was −$3 662.69 on income of $285 736.59 during the quarter on which data were available (the quarter coinciding with attendance at the assessment centre). After checking with the firm to see if a coding error had occurred, we learned that the firm was not making a profit from most of its 900 stores. In contrast, the approximately 1000 franchise stores (not under corporate ownership) were generating healthy profits. Corporate human resource professionals indicated that the company was considering selling all retail outlets within the next 18 months if financial performance did not improve. They candidly speculated that corporate bureaucracy prevented the stores from being responsive enough to local economic changes. Regardless, it is interesting to speculate that the correlation of overall assessment rating with store profit may be due to the absence of systematic and/or random error variation added by other contributors to

profit variance. In other words, a manager's influence on profit may have been detected because other influences, normally found in non-bureaucratic firms, were held constant within this single firm.

Despite the findings reported above, a number of deficiencies and alternative explanations in Study 1 prevent drawing strong conclusions for theory and practice. First, there was no manipulation check for assessor training. Regardless of the content of assessor training and their comments to the contrary, assessors may have been making unconscious attributions about candidates' latent skills and abilities in arriving at the dimensional ratings. The fact that this assessment centre used traditional dimensional labels like 'initiative' and 'judgement' that can easily be thought of as either skills and abilities or job requirements makes this a real possibility.

Second, assessors may have been able to generate construct valid ratings of both candidate traits and future performance on role requirements in the target job. When assessors are trained to rate only one type of construct (trait-oriented in a traditional assessment centre versus job-oriented in Study 1), the presence of task and role requirement information (Byham, 1977) may cause enough confusion that subsequent ratings do not yield expected evidence of convergent and discriminant validity. If assessors were trained to make both kinds of ratings, evidence of construct validity for both trait and job-oriented dimensions may be present. As assessors were only trained to provide performance consistency/role congruency ratings, this possibility could not be examined in Study 1.

Finally, assessors may simply not be capable of generating reliable and valid ratings of trait-oriented dimensions (as per the concerns voiced by assessors in Study 1). Instead, assessors may be using task-specific constructs to guide their evaluations, causing studies of post-exercise ratings to yield poor evidence of convergent and discriminant validity.

Hence, a better test of Klimoski & Brickner's (1987) performance consistency or Russell's (1987) role congruency explanation would involve an assessment centre in which ratings are made on *both* traditional trait-based dimensions and task-based dimensions. If assessors are only capable of making valid assessment of task-based dimensions, the task-based dimensions should yield criterion-related validities that are comparable to those found in traditional assessment centres, while trait-based dimensional ratings should yield lower, perhaps non-significant, criterion-related validities. Serendipitously, a field test of these hypotheses became possible in another division of the firm described in Study 1, using an assessment centre process developed 15 years earlier, and made available to the authors for research purposes.

STUDY 2

Method

Subjects

A sample of 172 current unit managers participated in this study. Unit managers were the lowest level supervisory position in the world-wide manufacturing facilities of the firm described in Study 1. Participants were randomly chosen from US facilities for participation in the validation study. Participants' immediate supervisors provided performance appraisal ratings. Assessors consisted of plant-level management personnel drawn from different manufacturing facilities who were unfamiliar with the candidates being assessed.

Job analysis and performance appraisal instrument

Structured interviews were conducted with approximately 50 incumbents and their supervisors to identify behavioural and task requirements of the unit manager position. Lists of behavioural and task requirements were derived from these interviews and circulated to upper-level plant management personnel for comment. A final list of 10 tasks with behavioural requirements was used to develop a performance appraisal instrument. Each task label (e.g. managing quality) was followed by a written definition of the task and a list of the specific behaviours needed for successful task completion. Participants' supervisors were asked to rate their subordinates' performance on each of the 10 tasks identified from the job analysis, using the task definitions and behavioural requirements for each task in arriving at a rating on a five-point scale (where 5 is high). Additionally, after providing ratings on the 10 tasks, superiors were asked to provide an overall performance rating. For 120 of the participating unit managers, performance appraisals were obtained from their immediate supervisor and a senior plant manager who had observed the participants' job performance. Participants' supervisors and senior plant managers were unaware of subordinates' assessment centre ratings.

Assessment procedures

The assessment centre procedures had been designed 15 years earlier for a 'foreman' position that had been the previous entry-level supervisory position in the manufacturing plants. While the current authors were directly involved in design of the performance appraisal instrument, we had no initial knowledge of the assessment centre content or procedures. A recent reorganization of the facilities had caused maintenance responsibilities to be added to the job description, along with a change in the job title to unit manager. The firm had entered into a consent decree with the US Justice Department over alleged irregularities in selection systems used in non-manufacturing (and non-retail) divisions. The consent decree obligated the firm to perform a criterion-related validity study whenever it implemented or revised selection procedure (in this case, due to a change in job requirements) that affected a large number of employees. The authors agreed to develop a performance appraisal instrument for purposes of re-validating the assessment procedure, resulting in the findings reported below.

When the assessment centre procedures were developed for the foreman position in the mid-1970s, a job analysis was performed, assessment dimensions identified, exercises and assessor training procedures developed, and a criterion-related validity study performed. All of these efforts were performed by a third-party vendor organization. Discussions with principals of this consulting firm indicated that task-related information from the job analysis was used to construct assessment centre exercises in much the same manner as described in Study 1, i.e. subject matter experts in the firm reviewed and modified exercise content to make it similar in scope and difficulty to tasks found in the target job. Four exercises (completed in one day) consisted of an in-basket, a manufacturing exercise, a management problems exercise, and an interview simulation. Assessor training consisted of two days of lecture, role playing, practice rating sessions, and feedback. Assessor ratings and consensus discussion took place in the same manner as described in Study 1.

Results from the original criterion-related validity study (conducted when the centre was first implemented) could not be located. However, conversations with the principals of the vendor organization that performed the criterion-related validity study indicated that concurrent validities obtained between the overall assessment rating (OAR) and incumbents' performance appraisal ratings were around .35. Hence, it can be initially assumed that task requirements in the target job were adequately represented in the assessment exercises with the possible exception of tasks related to maintenance responsibilities.

To our surprise, after running the sample of current store managers through the assessment centre and obtaining performance appraisal ratings, we learned that the centre had 'evolved' away from its original design. In contrast to assessment centre procedures typically found in industry, the firm had (on its own) modified traditional assessment procedures. Specifically, assessors were first asked to produce both ratings of typical person characteristic, skill, and ability-oriented assessment centre dimensions identified from the job analysis (e.g. 27 dimensions including initiative, listening skills, forcefulness, etc.) and then asked to rate seven dimensions of task performance found in the target job. The seven dimensions of task performance derived 15 years ago, when the centre was originally developed, paralleled the task statements derived from the current job analysis almost exactly as can be seen in Table 4 (these tasks were also used in construction of the assessment exercises). The clear similarity between these two lists is additional evidence of assessment centre content validity with respect to the unit manager position.

Table 4. Tasks used in the performance appraisal instrument and assessment centre

Assessment centre tasks (TP1–7)	Performance appraisal tasks
TP1. Working with subordinates	1. Orientation, training and development of workers 2. Working with subordinates
TP2. Maintaining efficient quality production	3. Maintaining quality 4. Maintaining efficient production
TP3. Organizing work of subordinates	5. Scheduling labour
TP4. Maintaining safe/clean work areas	6. Maintaining safe work conditions
TP5. Maintaining equipment and machinery	7. Maintaining machinery and solving technical problems
TP6. Handling routine supervisor responsibilities	8. Housekeeping 9. Time record keeping
TP7. Planning and scheduling	10. Planning production

Assessors were instructed to arrive at task-based ratings of how well they thought a candidate would perform that aspect of the target job based on the behaviours observed in those tasks within each assessment exercise. No one could be located in the organization who knew why this feature had evolved in the assessment design. The 27 person characteristic, skill, and ability-(trait) based ratings were grouped into categories labelled decision making ability, communication skills, managerial skills, personal characteristics and general abilities in the original assessment centre design and are described in Table 5.

All dimensional ratings were on a 1–5-point scale (where 5 is high), while the overall ratings were made on a 1–4-point scale (where 4 is high). Dimensional ratings were made after all exercises were completed using the AT&T prototype. Assessors did not know the participants and were unaware of participants' performance appraisal ratings. The three assessors observing and evaluating each candidate were than asked to arrive at an overall rating based on the seven task ratings using traditional consensus discussion procedures. A sum of all the ratings made on person characteristic-based assessment centre ratings were derived prior to arriving at an overall rating based on all information gathered in the assessment centre. Hence, three overall ratings were derived: (1) assessors' clinical combination of the seven task ratings; (2) a simple sum of the 27 trait-based dimensional ratings; and (3) assessors' clinical combination of (1) and (2).

Analyses

As the major purpose of this study is to examine the differential criterion-related validity of trait-based and task-based dimensional ratings, analyses focused on the psychometric characteristics of the criterion performance measure and the concurrent validity correlation coefficients between assessment centre ratings and the performance measure. The performance appraisal instrument was factor analysed to determine whether any groups of task requirements could be identified (the job analysis yielded no *a priori* groupings of task requirements). Any resultant factors or groupings of tasks were examined for internal consistency reliability. Simple correlations between assessment centre dimensions and resultant criterion measures were derived to evaluate whether task-based and trait-based assessment ratings were differently related to performance appraisal ratings.

Table 5. Person characteristic-based assessment dimensions

I. Decision-making ability
   1. Problem analysis
   2. Initiative
   3. Choice of alternatives
   4. Decisiveness
   5. Judgement

II. Communication skills
   6. Oral communication
   7. Written communication
   8. Verbal ability
   9. Listening skill

III. Managerial skills
   10. Willingness to be a leader
   11. Ability to delegate
   12. Achievement motivation
   13. Flexibility
   14. Persuasiveness
   15. Work situation awareness
   16. Job involvement
   17. Human relations skills
   18. Organizing and planning

IV. Person characteristics
   19. Personal impact
   20. Self-confidence
   21. Tolerance of stress
   22. Forcefulness
   23. Tenacity
   24. Physical energy
   25. Creativity

V. General abilities
   26. Mechanical ability
   27. Numerical ability

Results

Reliability of performance appraisal data

Common factor analysis of the performance appraisal ratings yielded a clean, one-factor solution. Internal consistency reliability for all 10 performance appraisal ratings was .88 while inter-rater reliability ($N = 120$ immediate supervisor–plant manager pairs) was .82. Ratings were averaged for the 120 subjects who had more than one performance appraisal rating.

Psychometric properties of assessment centre ratings

Common factor analysis with varimax rotation of dimensional ratings resulted in a clean four-factor solution explaining 85 per cent of the total variance. The seven task-based rat-

ings loaded on the first factor (average loading of .945 and an eigenvalue of 6.26). The 27 trait-based dimensions loaded cleanly on the remaining three factors (loading rule of .55 on the major factor and .40 on minor factors—average loadings on major and minor factors were .78 and .33 respectively). Dimensional ratings made on the groupings of decision-making ability, communication skills and managerial skills loaded on the second factor, ratings made on personal characteristics dimensions loaded on factor three, while ratings from general abilities loaded on factor four (see Table 5 for dimensional labels in each category). Except for the presence of the task-based ratings in factor one, these loadings are very similar to those reported by Russell (1985) and Sackett & Hakel (1979). Regardless, it appeared that a common set of behavioural observations yielded task-based ratings loading on a distinct factor (factor 1) from ratings of person characteristics, skills and abilities (factors 2, 3 and 4). This result is consistent with the trait versus task distinction assessors were trained to make when arriving at these ratings.

An R^2 of .88 was obtained when the overall rating derived from the seven task-based dimensional ratings was regressed on the task-based ratings. As reported in Study 1 and consistent with prior research (Russell, 1985; Sackett & Hakel, 1979), it appears that the majority of variance in the clinically derived overall rating is explained by the dimensional ratings the assessors were trained to consider.

Criterion-related validity analysis

All simple correlations between predictors (27 trait ratings + 1 sum of trait ratings + 7 task-based ratings + 1 overall task-based rating + 1 overall rating that combines trait-based and task-based information = 37 predictors) and criteria (10 performance appraisal ratings + 1 simple sum of the 10 ratings + 1 global performance rating = 12 criterion measures) yields a 49 × 49 correlation matrix that is available from the authors. Selected correlations reported in Table 6 indicate that task-based assessment centre ratings are consistently related to the overall performance rating, labelled PERFORM, and the sum of the 10 performance dimension ratings, labelled PASUM (the average correlation with PERFORM and PASUM was .22, $N = 14$, SD = .04). The simple correlations between the overall task-based assessment rating (OARTASK) and performance ratings are .28 ($p < .0001$) for PERFORM and .33 ($p < .0001$) for PASUM. Further, a simple sum of the task-based ratings (TASKSUM) is also highly correlated with both performance measures (.27, $p < .0001$), with PERFORM and .32, $p < .0001$, with PASUM). These are typical of the criterion-related validities reported by Gaugler *et al.*'s (1987) meta-analysis.

In contrast, correlations between the trait-based assessment ratings and performance ratings were significantly lower and less consistent (average correlation with PERFORM and PASUM was .18, $N = 54$, SD = .09). Further, the correlation between the sum of the trait-based ratings and PERFORM was .138 (non-significant), while the correlation with PASUM was .143 (non-significant), less than half as large as the correlations reported above for task-based overall rating. Correlations between unit-weighted scale scores (i.e. factors 2–4 from the factor analysis conducted on all assessment dimensional ratings) and performance measures were all non-significant. Note that because there was no clinically derived OAR from trait-based ratings, the comparison to task-based OAR ratings cannot be complete.

Table 6. Descriptive statistics and correlations among assessment centre and performance appraisal ratings ($N = 172$)

Variables[a]	Mean	SD	TP1	TP2	TP3	TP4	TP5	TP6	TP7	OARTASK
TP1-Working with subordinates	3.26	.70	–							
TP2-Efficient quality production	3.27	.68	.82	–						
TP3-Organizing subordinate work	3.16	.73	.77	.77	–					
TP4-Clean/safe work areas	3.28	.66	.84	.87	.78	–				
TP5-Equipment maintenance	3.28	.70	.71	.75	.70	.76	–			
TP6-Routine supervisory responsibility	3.22	.70	.69	.71	.68	.73	.67	–		
TP7-Planning and scheduling	3.11	.72	.77	.81	.87	.79	.74	.76	–	
OARTASK	3.23	.69	.87	.90	.83	.88	.79	.77	.85	–
PC1-Problem analysis	3.23	.68	.73	.76	.79	.75	.70	.69	.78	.79
PC2-Initiative	3.24	.75	.69	.72	.65	.71	.63	.60	.69	.71
PC3-Choice of alternatives	3.01	.68	.66	.70	.80	.68	.62	.62	.75	.72
PC4-Decisiveness	3.23	.75	.74	.76	.77	.80	.65	.64	.74	.80
PC5-Judgement	3.16	.65	.65	.65	.70	.65	.65	.56	.66	.70
PC6-Oral communication	3.63	.62	.68	.70	.67	.68	.60	.61	.62	.68
PC7-Written communication	3.52	.80	.44	.42	.46	.40	.35	.51	.42	.43
PC8-Verbal ability	3.41	.67	.65	.67	.68	.65	.63	.67	.68	.73
PC9-Listening skill	3.39	.59	.62	.62	.65	.56	.54	.61	.64	.64
PC10-Willingness to lead	3.22	.80	.74	.78	.74	.78	.68	.69	.77	.77
PC11-Ability to delegate	3.13	.83	.65	.72	.69	.70	.65	.65	.74	.70
PC12-Achievement motivation	3.37	.68	.68	.74	.63	.71	.65	.60	.65	.70
PC13-Flexibility	3.41	.66	.52	.51	.52	.52	.48	.43	.49	.51
PC14-Persuasiveness	3.17	.80	.71	.74	.68	.75	.65	.65	.69	.76
PC15-Situation awareness	3.28	.72	.68	.73	.68	.72	.67	.64	.69	.72
PC16-Job involvement	3.45	.68	.69	.73	.69	.73	.65	.63	.67	.72
PC17-Human relations skill	3.54	.58	.53	.47	.47	.46	.41	.44	.44	.49
PC18-Organizing and planning	3.10	.77	.70	.76	.82	.74	.71	.70	.82	.79
PC19-Personal impact	3.51	.67	.67	.66	.59	.67	.59	.58	.60	.66
PC20-Self-confidence	3.38	.85	.73	.75	.77	.76	.64	.66	.72	.77
PC21-Tolerance of stress	3.45	.60	.67	.62	.59	.60	.51	.54	.55	.64
PC22-Forcefulness	3.17	.83	.68	.70	.68	.74	.63	.61	.68	.69

Table 6 *(cont.)*

PC23-Tenacity	3.41	.73	.62	.61	.64	.63	.55	.57	.58	.61
PC24-Physical energy	3.45	.67	.61	.71	.59	.64	.60	.56	.60	.68
PC25-Creativity	3.17	.69	.68	.70	.76	.72	.68	.58	.72	.73
PC26-Mechanical ability	3.21	1.14	.29	.31	.31	.34	.56	.28	.33	.36
PC27-Numerical ability	2.04	1.24	.36	.35	.35	.38	.42	.54	.42	.39
SUMPC	264.20	42.88	.51	.48	.49	.47	.42	.39	.49	.48
PA1-Planning production	3.28	.76	.29	.30	.15	.22	.16	.17	.21	.25
PA2-Scheduling labour	3.29	.80	.24	.21	.16	.20	.17	.17	.16	.21
PA3-Maintaining quality	3.04	.78	.26	.28	.18	.23	.17	.20	.18	.24
PA4-Efficient production	3.13	.78	.27	.31	.20	.19	.22	.21	.19	.25
PA5-Maintaining machinery	3.11	.78	.30	.31	.26	.29	.35	.27	.26	.33
PA6-Housekeeping	2.73	.68	.21	.14	.18	.13	.14	.15	.14	.23
PA7-Record keeping	3.24	.75	.25	.27	.19	.21	.20	.26	.19	.26
PA8-Safety	3.28	.66	.19	.19	.12	.20	.22	.20	.18	.22
PA9-Development of workers	2.96	.70	.32	.32	.29	.30	.29	.32	.29	.32
PA10-Working with subordinates	3.20	.76	.21	.28	.16	.20	.20	.22	.18	.24
PASUM	62.55	11.40	.34	.34	.24	.28	.27	.27	.25	.33
Overall PERFORMance rating	3.21	.66	.27	.31	.19	.22	.23	.21	.24	.28

Variables[a]	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
PC1-Problem analysis	–													
PC2-Initiative	.62	–												
PC3-Choice of alternatives	.64	.63	–											
PC4-Decisiveness	.71	.69	.67	–										
PC5-Judgement	.67	.56	.67	.63	–									
PC6-Oral communication	.64	.64	.58	.65	.55	–								
PC7-Written communication	.47	.39	.47	.41	.38	.40	–							
PC8-Verbal ability	.68	.57	.65	.68	.64	.58	.46	–						
PC9-Listening skill	.59	.54	.64	.57	.51	.59	.51	.58	–					
PC10-Willingness to lead	.69	.72	.63	.75	.58	.68	.45	.61	.55	–				
PC11-Ability to delegate	.66	.68	.73	.67	.59	.59	.49	.60	.59	.73	–			

Table 6 (*cont.*)

Variables[a]	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
PC12-Achievement motivation	.66	.64	.60	.66	.63	.68	.33	.59	.48	.66	.62	–		
PC13-Flexibility	.50	.47	.54	.43	.57	.52	.32	.50	.45	.38	.45	.61	–	
PC14-Persuasiveness	.66	.75	.57	.73	.60	.65	.37	.64	.50	.75	.64	.69	.48	–
PC15-Situation awareness	.64	.61	.63	.66	.65	.66	.29	.60	.59	.69	.61	.71	.52	.66
PC16-Job involvement	.65	.64	.66	.63	.60	.62	.48	.61	.53	.62	.61	.70	.54	.64
PC17-Human relations skill	.43	.45	.48	.42	.52	.48	.25	.39	.48	.38	.44	.49	.56	.43
PC18-Organizing and planning	.74	.59	.73	.70	.67	.59	.45	.62	.61	.70	.69	.64	.53	.64
PC19-Personal impact	.56	.62	.55	.61	.56	.69	.36	.56	.55	.65	.56	.59	.47	.66
PC20-Self-confidence	.67	.70	.65	.75	.62	.71	.40	.65	.56	.81	.67	.64	.38	.78
PC21-Tolerance of stress	.59	.48	.50	.53	.48	.62	.28	.53	.50	.53	.49	.58	.51	.56
PC22-Forcefulness	.64	.68	.57	.71	.55	.67	.40	.60	.45	.78	.65	.68	.38	.75
PC23-Tenacity	.60	.59	.56	.59	.58	.61	.41	.60	.47	.59	.52	.65	.61	.61
PC24-Physical energy	.57	.54	.58	.59	.53	.63	.37	.61	.49	.61	.56	.66	.46	.64
PC25-Creativity	.69	.61	.67	.67	.73	.60	.40	.58	.50	.66	.57	.62	.50	.62
PC26-Mechanical ability	.36	.29	.27	.24	.29	.23	.12	.26	.30	.27	.25	.27	.28	.32
PC27-Numerical ability	.35	.25	.31	.23	.29	.39	.29	.33	.39	.30	.34	.36	.33	.34
SUMPC	.50	.46	.45	.45	.50	.48	.25	.44	.43	.51	.48	.50	.42	.46
PA1-Planning production	.20	.37	.15	.19	.12	.33	.20	.20	.19	.27	.19	.22	.15	.19
PA2-Scheduling labour	.16	.25	.15	.22	.09	.32	.14	.21	.12	.25	.16	.14	.10	.15
PA3-Maintaining quality	.19	.20	.12	.21	.11	.35	.22	.18	.15	.27	.19.	20	.13	.18
PA4-Efficient production	.19	.15	.17	.23	.18	.29	.16	.22	.18	.27	.19	.28	.12	.23
PA5-Machinery maintenance	.27	.26	.21	.22	.21	.32	.17	.27	.22	.33	.21	.25	.22	.25
PA6-Housekeeping	.13	.23	.12	.15	.04	.19	.12	.20	.11	.15	.07	.13	.05	.10
PA7-Record keeping	.19	.14	.13	.14	.14	.30	.14	.17	.14	.18	.13	.21	.18	.18
PA8-Safety	.08	.20	.11	.16	.02	.23	.04	.14	.05	.22	.06	.17	.02	.25
PA9-Worker development	.29	.24	.26	.26	.23	.42	.21	.30	.23	.32	.28	.30	.26	.25
PA10-Works with subordinates	.14	.22	.14	.26	.14	.32	.14	.16	.09	.29	.16	.29	.10	.18
Sum of PA1 to PA10	.24	.17	.20	.26	.16	.41	.20	.27	.20	.34	.22	.28	.17	.27
Overall PERFORMance rating	.21	.22	.15	.18	.11	.32	.14	.20	.13	.26	.17	.26	.14	.39

Table 6 (*cont.*)

Variables[a]	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22	PC23	PC24	PC25	PC26	PC27	SUMPC
PC15-Situation awareness	–													
PC16-Job involvement	.68	–												
PC17-Human relations skill	.47	.44	–											
PC18-Organization and planning	.67	.67	.41	–										
PC19-Personal impact	.62	.58	.58	.57	–									
PC20-Self-confidence	.67	.68	.43	.68	.69	–								
PC21-Stress tolerance	.55	.53	.55	.52	.60	.61	–							
PC22-Forcefulness	.63	.66	.33	.63	.61	.76	.51	–						
PC23-Tenacity	.59	.66	.45	.56	.57	.60	.55	.63	–					
PC24-Physical energy	.65	.63	.43	.51	.58	.65	.57	.60	.56	–				
PC25-Creativity	.65	.65	.46	.69	.58	.67	.55	.67	.62	.58	–			
PC26-Mechanical ability	.30	.25	.20	.35	.22	.16	.25	.24	.22	.25	.37	–		
PC27-Numerical ability	.36	.28	.25	.42	.29	.23	.31	.20	.31	.26	.31	.45	–	
SUMPC	.51	.46	.42	.49	.43	.50	.40	.42	.47	.43	.47	.18	.27	–
PA1-Planning production	.21	.12	.14	.13	.24	.19	.17	.21	.13	.22	.09	.14	.16	.21
PA2-Scheduling labour	.12	.08	.06	.11	.22	.22	.11	.17	.07	.16	.05	.16	.19	.13
PA3-Maintaining quality	.21	.19	.07	.11.	.22	.20	.17	.19	.13	.19	.09	.07	.16	.13
PA4-Efficient production	.22	.18	.15	.17	.21	.29	.23	.18	.18	.22	.15	.20	.18	.16
PA5-Maintaining machinery	.29	.26	.16	.27	.27	.29	.19	.28	.15	.18	.23	.38	.26	.12
PA6-Housekeeping	.05	.07	.07	.08	.12	.13	.15	.06	.04	.11	.05	.09	.12	.09
PA7-Record keeping	.22	.17	.18	.21	.25	.17	.22.	.16	.20	.13	.17	.12	.28	.03
PA8-Safety	.10	.12	−.00	.10	.17	.17	.11	.22	.04	.15	.08	.19	.15	−.03
PA9-Worker development	.25	.28	.12	.26	.27	.27	.33	.27	.26	.23	.21	.21	.29	.17
PA10-Works with subordinates	.21	.14	.07	.19	.19	.26	.20	.23	.23	.20	.14	.11	.18	.08
Sum of PA1 to PA10	.25	.22	.12	.21	.29	.30	.25	.26	.19	.24	.17	.21	.24	.14

Table 6 (*cont.*)

Variables[a]	PA1	PA2	PA3	PA4	PA5	PA6	PA7	PA8	PA9	PA10	PASUM	PERFORM
PA1-Planning production	–											
PA2-Scheduling labour	.78	–										
PA3-Maintaining quality	.69	.58	–									
PA4-Efficient production	.71	.67	.64	–								
PA5-Machinery maintenance	.57	.63	.54	.65	–							
PA6-Housekeeping	.53	.56	.52	.44	.47	–						
PA7-Record keeping	.58	.50	.54	.55	.45	.51	–					
PA8-Safety	.46	.44	.51	.46	.48	.56	.46	–				
PA9-Worker development	.66	.60	.65	.53	.53	.53	.59	.49	–			
PA10-Works with subordinates	.69	.66	.58	.64	.48	.41	.62	.47	.63	–		
Sum of PA1 to PA10	.86	.82	.80	.82	.75	.69	.75	.67	.79	.80	–	
Overall PERFORMance rating	.84	.75	.74	.77	.67	.58	.68	.58	.72	.76	.81	–

[a] All correlations >.15 are significantly different from zero in a two-tailed test at $p < .05$; All correlations >.18 are significantly different from zero in a two-tailed test at $p < .01$.

Finally, the correlation between the overall assessment rating (OAR—assessors' final consensus judgement based on all trait and task-based ratings) and the overall performance ratings (PERFORM) was .18, while the correlation with the sum of the 10 performance appraisal ratings (PASUM) was .15. Given the criterion-related validity reported above for task-based ratings (i.e. the .27–.33 range), it appears that trait-based information dilutes the criterion-related validity generated from task-based ratings when combined through consensus discussion to arrive at a final OAR.

Discussion

The results reported above clearly support the performance consistency/role congruency explanation of assessment centre construct validity. Criterion-related validities were higher for the task-based assessment dimensions and displayed more consistency (less variation) in their correlations with performance criteria. Indeed, when assessors were asked to consider trait-based ratings in combination with the task-based ratings, criterion-related validity was reduced. If this were the first study to have reported results suggesting a performance consistency/role congruency explanation, we would be much more cautious in our conclusion. However, when asked to provide role congruency ratings in Study 1, criterion-related validities consistent with those reported in traditional assessment centres were observed. In Study 2, when assessors were asked to provide *both* trait-based and task-based ratings, trait-based ratings demonstrated substantially lower criterion-related validity while task-based ratings exhibited the customarily high criterion-related validities reported in the literature. When assessors have been asked to provide trait-based ratings alone, evidence of criterion-related validity in the absence of construct validity (Bycio *et al.*, 1987; Neidig *et al.*, 1979; Russell, 1987; Sackett & Dreher, 1984; Shore *et al.*, 1990; Turnage & Muchinsky, 1982) suggests that typical assessors are *not* providing trait-based ratings. Instead, in light of Study 1 results, assessors in traditional assessment centres may actually provide criterion valid task-based ratings that clearly should not (and do not) exhibit evidence of trait-based construct validity.

Of course, the conventional explanation for these results is that there was some fatal flaw in the assessment design or assessor training that prevented trait-based ratings from demonstrating criterion-related validity (e.g. common halo in both assessment and performance appraisal ratings). While the authors have a great deal of prior experience developing and implementing traditional trait-oriented assessment centres that do demonstrate criterion-related validity, this possibility can only be ruled out through independent replication. Future research replicating Study 2 and extending it to the post-exercise assessment centre prototype (to examine convergent and discriminant validity) is needed.

Our only concern in rejecting the trait-based explanation of assessment centre construct validity is that these results will need to be examined in the small proportion of assessment centres operating in non-selection environments. These centres attempt to assess training and development needs (many staffed by doctoral level psychologists as assessors) and hence may indeed yield construct valid trait-based ratings. In partial support of this speculation, Gaugler *et al.* (1987) found that the presence of psychologists as assessors moderated (increased) the criterion-related validities reported in their meta-analysis.

Conclusion

It would appear that, in future assessment centres used to make management selection decisions, assessment centre architects should consider redesigning both exercise construction and dimension specifications to make them congruent with the performance consistency/role congruency explanation. Specifically, vast amount of assessor training time is currently taken up in repeated exposure to exercise role plays and discussions of why an observed behaviour is or is not representative of some skill or ability (typical assessor training programmes offered in industry and by vendors can last one week or more, e.g. Dugan, 1988). Further, the task-based procedure is amenable to content validity and synthetic validity arguments while traditionally designed assessment centres are not (Sackett, 1982). Criterion validity demonstrated here might be transported to other management positions that exhibit comparable role and performance requirements in the job analysis.

However, the exercises cannot be considered a job sample—these were simulation exercises that took place for one day in hotel rooms around the United States. This raises a number of interesting questions. For example, if a sample of actual job tasks and a set of exercise tasks have overlapping role requirements, how well will behaviours elicited in the exercises predict task performance on the job? Specifically, how deviant can the tasks involved in the exercises be from actual job requirements before dimensional ratings will not exhibit criterion validity? Will any set of tasks that permit candidates to demonstrate the necessary behaviours yield criterion valid role congruency ratings? Recent findings reported by Motowidlo, Dunnette & Carter (1990) suggest that 'low fidelity' simulations yield criterion-related validities comparable to those found in the assessment centre literature, though they do not explore constructs associated with behaviours in these simulations. Future work needs to examine relationships between job requirements (tasks), construct validity of behavioural role requirements as reflected in dimensional ratings, and exercise construction (high vs. low fidelity). Furthermore, this work needs to examine the impact of multiple assessment prototypes (AT&T versus post-exercise formats) on these relationships.

In sum, the performance consistency/role congruency explanation appears to be the most parsimonious explanation of prior efforts focused on determining why assessment centres demonstrate criterion-related validity. We find it intuitively appealing to view assessment centre exercises and simulations as structured ways of obtaining a sample of behaviours. Asking assessors to catalogue these behavioural observations into clusters that correspond with basic task and role requirements of the job does not need to be complicated with notions of person characteristics, skills and abilities. Assessor training and labels used to describe assessment dimensions should be modified accordingly.

Acknowledgements

We would like to thank James Sharf for helpful suggestions made throughout this project. Philip Bobko, Mike Campion and Adrienne Colella provided useful comments on an earlier version of the manuscript.

References

Bass, B. F. (1949). An analysis of the leaderless group discussion. *Journal of Applied Psychology*, 33, 527–533.

Bycio, P., Alveres K. M. & Hahn, J. (1987). Situation specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology*, 72, 463–474.

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

Byham, W. C. (1977). Assessor selection and training. In J. L. Moses & W. C. Byham (Eds), *Applying the Assessment Center Method*, pp. 89–127. New York: Pergamon.

Byham, W. C. (1980). Assessor selection and training. In J. L. Moses & W. C. Byham (Eds), *Applying the Assessment Center Method*, pp. 89–126. New York: Pergamon Press.

Cranny, C. J. & Doherty, M. E. (1988). Importance ratings in job analysis: Note on the misinterpretation of factor analyses. *Journal of Applied Psychology*, 73, 320–322.

Crooks, L. A. (1977). The selection and development of assessment center techniques. In J. L. Moses & W. A. Byham (Eds), *Applying the Assessment Center Method*, pp. 69–88. New York: Pergamon Press.

Dugan, B. (1988). Effects of assessor training on information use. *Journal of Applied Psychology*, 73, 743–748.

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

Hemphill, J. K., Griffiths, D. E. & Frederiksen N. (1962). *Administrative Performance and Personality: A Study of the Principal in a Simulated Elementary School.* New York: Teachers College Bureau of Publications, Columbia University.

Holmes, D. S. (1977). How and why assessment works. In J. L. Moses & W. C. Byham (Eds), *Applying the Assessment Center Method*, pp. 127–143. New York: Pergamon Press.

Howard, A. & Bray, D. W. (1988). *Managerial Lives in Transition: Advancing Age and Changing Times.* New York: Guilford Press.

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

Klimoski, R. J. & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, 40, 243–260.

Motowidlo, S. J., Dunnette, M. D. & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, 75, 640–647.

Neidig, R. D., Martin, J. C. & Yates, R. E. (1979). The contribution of exercise skill ratings to final assessment center evaluations. *Journal of Assessment Center Technology*, 2, 21–23.

Russell, C. J. (1985). Individual decision processes in an assessment center. *Journal of Applied Psychology*, 70, 737–746.

Russell, C. J. (1987). Person characteristic vs. role congruency explanations for assessment center ratings. *Academy of Management Journal*, 30, 817–826.

Sackett, P. R. (1982). A critical look at some common beliefs about assessment centers. *Public Personnel Management*, 11, 140–146.

Sackett, P. R. & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling findings. *Journal of Applied Psychology*, 67, 401–410.

Sackett, P. R. & Dreher, G. F. (1984). Situation specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. *Journal of Applied Psychology*, 69, 187–190.

Sackett, P. R. & Hakel, M. D. (1979). Temporal stability and individual differences in using assessment center information to form overall ratings. *Organizational Behavior and Human Performance*, 23, 120–137.

Sackett, P. R. & Harris, M. M. (1988). A further examination of the constructs underlying assessment center ratings. *Journal of Business and Psychology*, 3, 214–229.

Shore, T. H., Thornton, G. C. III & Shore, L. M. (1990). Constuct validity of two categories of assessment center dimension ratings. *Personnel Psychology*, 43, 101–116.

Standards for Ethical Considerations for Assessment Center Operations (1977). In J. L. Moses & W. C. Byham (Eds), *Applying the Assessment Center Method,* pp. 303–310. New York: Pergamon Press.

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

Turnage, J. J. & Muchinsky, P. M. (1982). Transituational variability in human performance within assessment centers. *Organizational Behavior and Human Performance*, 30, 174–200.

Received 18 January 1993; final version received 10 August 1994