European Journal of Psychological Assessment, Vol. 17, Issue 1, pp. 17–24

Appraisal Ratings, Halo, and Selection: A Study Using Sales Staff

Chris J. Jackson[1] and Adrian Furnham[2]

[1]University of Queensland, Australia, [2]University College London, UK

Keywords: Rating error, halo, appraisal, selection, sales

Summary: Supervisor ratings are useful criteria for the validation of selection instruments but may be limited because of the presence of rating errors, such as halo. This study set out to show that supervisor ratings which are high in halo remain successful criteria in selection. Following a thorough job analysis, a customer service questionnaire was designed to assess the potential of retail sales staff on three "orthogonal" subscales labelled *Dealing with people*, *Emotions and energy*, and *Solitary style*. These subscales were uncorrelated with supervisor ratings made about 8 weeks later. However, the supervisor ratings were correlated with an overall scale derived from the three scales of the customer service questionnaire. These results support the view that supervisor ratings generally consist of global impressions and suggest that these global impressions are useful measures of overall performance. This field study confirms laboratory results that halo does not necessarily reduce rating accuracy.

In performance appraisal, workers are often assessed on numerous rating scales that are designed to reflect their performance on job-related scales. Supervisor ratings are thought to be reasonable criteria for the validation of selection instruments (Hoffman, Nathan, & Holden, 1991; Nathan & Alexander, 1988), yet rating scales are also commonly believed to be prone to several different forms of rater bias, of which perhaps the best known is the halo rating error (Balzer & Sulsky, 1992; Landy & Farr, 1980). Halo has traditionally been seen as a type of rater error that occurs when a rater appraises others according to a global, overall impression; or in other words, when the observed correlation between rating scales is higher than could be expected (see Borman, 1977; Fisicaro, 1988; Lance, LaPointe, & Stewart, 1994). Halo is an observable end-product of a cognitive process in which systematic distortions result from implicit theories, person schemata or prototypes. Raters introduce such systematic distortions into the rating process for various reasons, the most important of which are thought to be low motivation, insufficient chances to observe the appraisee, and poor quality rating instruments (Cooper, 1981; Feldman, 1981; Landy & Farr, 1980).

The aim of using rating scales in performance appraisal is to achieve accurate assessments that reflect the true scores of the worker. This is because appraisal ratings are used to determine performance related pay, promotion and training needs. Since halo is seen as a rating error, the need to have highly accurate ratings suggests the need to keep halo to a minimum. However, a positive relationship between accuracy and invalid halo has been reported in the literature, suggesting an *increase* in invalid halo rating error is associated with an *increase* in rating accuracy (Kozlowski & Kirsch, 1987; Murphy & Balzer, 1986; Nathan & Tippins, 1990). This finding has been called the "halo-accuracy paradox" (Cooper, 1981; Jackson, 1996; Murphy & Cleveland, 1991). Jackson (1996) explains this finding by means of a model in which maximum achievable accuracy is not necessarily at the point where rating error is at its smallest. Two laboratory studies provided support for the hypothesis that accuracy can be positively related to the halo rating error. There is therefore a need for a field study to support

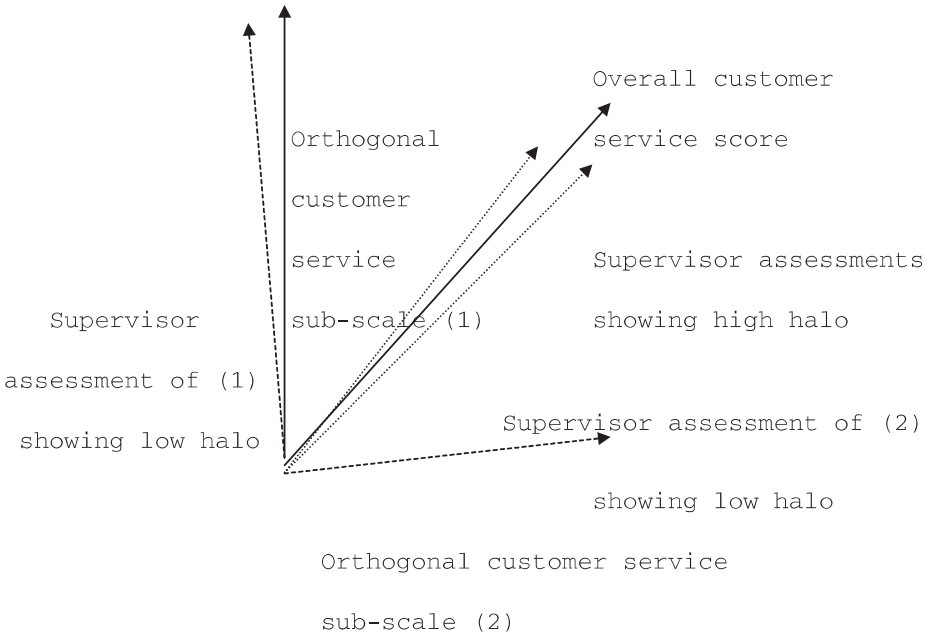

Figure 1. Graphical representation of the correlations between supervisor ratings with high levels of halo and scores from a customer service questionnaire. In this graphical representation, the solid-line vectors represent the scales of the customer service questionnaire and the short broken-line vectors represent the supervisor assessments. The smaller the angle between the vectors, the higher the correlation. Here, the overall scale of the customer service questionnaire can be expected to be more highly correlated to the supervisor assessments which contain the halo rating error than the sub-scales. However if the supervisor assessments contain only low levels of halo (long broken-line vectors) then the supervisor assessments could be more highly correlated with the sub-scales of the customer service questionnaire than the overall scale.

the general contention that ratings containing halo rating error can be reasonably accurate measures of staff performance. Such a field study could test the idea that some recent reviewers are incorrect when they suggest that the halo-accuracy paradox may be an artefact resulting from poor measurement techniques (Murphy & Balzer, 1989; Murphy & Cleveland, 1991).

One method for determining the accuracy of supervisor ratings in the field is to relate them to scores from a selection test that has been properly constructed; in this case a retail sales questionnaire designed to predict quality of customer service. Here we are investigating the usefulness of a "criterion" through its relationship with a selection instrument; an *inverse diagnostic method* that contrasts with the more standard statistical design of evaluating a selection instrument against a criterion.

In general, customer service questionnaires have relatively high validity against supervisor ratings. A quantitative review of the literature shows that the average validity of personality measures is 0.5, whereas the correlation against cognitive ability is low (Frei & McDaniel, 1998). Parasuraman, Zeithaml, and Berry (1985, 1988) have been particularly concerned with the development and refinement of an instrument for measuring customer perceptions of service. Parasuraman et al. (1985) identified ten determinants of service quality, some of which are dependent on the characteristics of the supplier of the service. These include *reliability* (consistency of performance and dependability), *responsiveness* (willingness or readiness to provide service), *competence* (required skills and knowledge to perform the service), *access* (approachability and ease of contact), *courtesy* (politeness, respect, consideration, and friendliness of contact personnel), *communication* (keeping the customers informed in language they can understand and listening to them), and *understanding* (knowing the customer and making the effort to understand the customer's needs). Carmen (1990) found that the instrument provided a useful "basic skeleton" for use across a broad spectrum of services, with its key dimensions exhibiting an important level of stability. However, the original dimensions identified by Parasuraman et al. (1985, 1988) are not completely generic. Thus, new dimensions may need to be investigated in some service contexts other than those used in this initial study. Carmen further recommended that when it is obvious to customers that multiple service functions are performed, the instrument, or any derivative, should be administered for each function separately.

A Customer Service Questionnaire has also been developed by a British Consultancy. This measure has three superfactors (Relations with people, Thinking Style, and Emotions and energy) and eleven factors. Furnham and Coveney (1996) related the questionnaire to the big five personality factors as defined by Costa and McCrae (1989). Ninety-two working adults in a major customer service business completed the Customer Service Questionnaire and the Costa and McCrae measure. The strongest correlations with the eleven customer service factors were positive for Extraversion (0.68), negative

for Neuroticism (–.67) and positive for Conscientiousness (0.55). The results provide some support for construct validity of the Customer Service Questionnaire.

Furnham (1994) reported that the subscales of this questionnaire were not highly correlated with each other (average $r = .17$) and was reasonably predictive of a number of dependent variables such as uncertified absenteeism, punctuality and letters of compliment. However this questionnaire was *not* predictive of supervisor ratings. Only three of the scales ("Approach to organizing," "Need for results," and "Need for social approval") were related to 50% or more of the supervisor assessments.

The aim of this paper is to show that high levels of halo in supervisor assessments do not reduce the level of accuracy of these assessments *at the overall level of analysis*. The specific hypothesis that we tested was: Supervisor ratings are significantly correlated (i. e., are useful criteria) with an overall assessment from a customer service questionnaire even when supervisor assessments are not useful criteria in the prediction of relatively orthogonal personality subscales from a customer service questionnaire.

Why might supervisor ratings predict overall scores, but not orthogonal personality subscale scores from a selection test? Because of the halo rating error, it seems likely that supervisor ratings on individual components of the job will be global impressions. All other things being equal, global appraisals are then more likely to be correlated with overall scores derived from a customer service questionnaire. In contrast, it is much less likely that global supervisor ratings will be correlated with orthogonal subscales of the questionnaire. Figure 1 provides a graphical representation of why this is the case. The solid-line vectors represent the subscales of the customer service questionnaire. Given that the subscales are orthogonal to each other (shown graphically as being at right angles to each other), the subscales are not highly correlated with the supervisor assessments which are shown by the short broken-line vectors. In fact, only the overall customer service score is highly correlated with the supervisor assessments (as shown by the small angle between the two vectors). Only if the supervisor assessments do not contain the halo rating error (as shown by the long broken-line vectors) are they highly correlated with the customer service subscales. Such results will confirm the prediction of Jackson (1996) that halo does not necessarily have large negative effects on accuracy, show the usefulness of understanding halo in the field and show that supervisor ratings are useful criteria even in the presence of the halo rating error.

Method

Participants

We studied 110 temporary retail sales staff employed by a large high street retail store on Oxford Street, London, UK, during the Christmas period. This store mainly sells clothes. Participants were instructed to complete the questionnaire as if they were applying for the job.

Development of Questionnaire

The retail sales questionnaire was designed after extensive organizational and job analysis. This involved understanding the culture in which people are required to operate, the content of the job or jobs performed in terms of the task and activities carried out by job holders, and the skills required to perform these tasks successfully and to operate within the climate.

The job analysis was conducted in a cross-section of three shops from three areas with a representative sample of staff. In each shop the Store Manager as well as newest, best, and "least good" sales staff were interviewed. Each interviewee was reassured that the information derived would in no way be harmful to their career, and that information would be kept anonymous. Furthermore, every effort was made to both explain the nature of the project as well as put the interviewee at ease. Each manager or sales consultant was interviewed about the content of their jobs in terms of the day-to-day activities

Table 1. Brief description of the contents of the Retail Sales Questionnaire as derived by job analysis.

Solitary:	High scoring sales people enjoy spending time on their own
Adaptable/Flexible:	High scoring sales staff are open to rational comment and are tolerant of uncertainty
Assertiveness:	High scorers are the centre of attention
Active:	High scoring sales staff satisfy their own needs and the needs of others
Stand up for one-self:	High scorers are determined and fixed in their views
Poor locus of control:	Sales staff feel guilt when things go wrong and let others take the credit for success
Satisfaction:	High scorers enjoy the job.
Work motivation:	Sales staff who are high scorers are keen and enthusiastic.
Emotional instability:	High scorers are moody, depressed and inconsistent in their approach to work, colleagues and customers
Customer orientation:	High scorers put the customer first and wish to provide a first class service.

Table 2. Correlations between the subscales and the content areas of the customer service questionnaire. This table shows the correlations between the original content areas that the questionnaire was designed to cover and the three subscales that represented these content areas. The number of items that covered each of the content areas is shown in brackets.

	Items covering each content areas	Dealing with people	Emotions and energy	Solitary style
Solitary	(18)	–.09	.09	.43**
Adaptable/Flexible	(16)	.28*	.26*	.06
Assertiveness	(5)	.64**	.57**	–.05
Active	(8)	.62**	.60**	.22
Stand up for oneself	(5)	.45**	.54**	.02
Poor attributional style	(15)	.12	.51**	.45**
Satisfaction	(15)	.86**	.53**	.18
Work motivation	(17)	.64**	.40**	.20
Emotional instability	(12)	.58**	.60**	.23*
Customer orientation	(21)	.24*	–.09	.11

*$p < .05$, **$p < .01$

Table 3. Means, standard deviations, internal consistency, and correlations between the scales developed from the customer service questionnaire.

Scale	Mean	*SD*	No. of items	α	D	E	S
Dealing with people (D)	203.7	20.2	63	.94			
Emotions and Energy (E)	88.7	15.7	46	.91	–.52**		
Solitary style (S)	38.5	6.5	23	.68	–.05	.17	
Overall customer service scale	65.5	8.4	31	.78	.58**	–.77**	–.33*

Each scale of the customer service questionnaire was rated using a four-point scale (4 = strongly agree, 3 = agree, 2 = disagree, 1 = strongly disagree). The supervisor rating was the sum of 9 scales used to assess the competence of retail sales staff. Each scale was rated using a four-point scale (3 = exceptional, 2 = highly effective, 1 = effective contribution, 0 = below standard). The correlations between the nine scales are shown in Table 4.

performed and their roles and responsibilities. The interviews were structured using the "critical incident" technique (Flanagan, 1954). The interview with the store manager lasted about 50 minutes, whereas the interview with the sales consultant lasted 20–35 minutes depending how expressive and knowledgeable the person was concerning their job. The content areas derived from the job analysis are shown in Table 1, and the number of items in the final questionnaire that reflected each subscale are shown in Table 2. It should be noted that some of these scales reflect the prevalent poor attributional style and poor culture found at this level in this organization.

By means of a scree plot, principal components analysis with varimax rotation indicated that three scales provided an effective summary of the ten content areas (explaining 50% of the variation in the dataset). These subscales were then further refined by means of item analysis and labelled *Dealing firmly with people* (D), *High emotions and energy* (E), and *Solitary style* (S) to reflect both the content areas and items that comprised these scales. The subscales were named so that high scorers were similar to the names of the subscales. The correlations between the subscales and the content areas are shown in Table 2. All the content areas have at least one reasonably high correlation with one of the orthogonal subscales and the correlations are in the expected direction.

The average correlation between the scales, after *r–z* transformation, was $r = 0.23$, suggesting a reasonably orthogonal relationship on average between the "personality" subscales (the actual correlations are shown in Table 3 and show that the only significant correlation between the three scales was between D and E). Item analysis was then used to refine and reduce the three scales to one overall scale. The overall scale was strongly correlated with each of the three original subscales (D, $r = .58$, $p < .01$; E, $r = .77$, $p < .01$; S, $r = .33$, $p < .05$). The three subscales consisted of between 23 and 63 items and the overall scale consisted of 31 items (as shown in Table 3). Means and standard deviations of the scales developed for use in the retail sales questionnaire and means

Table 4. Intercorrelations between the supervisor ratings across nine facets of the job.

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(I)
(b)	.12								
(c)	.52**	.52**							
(d)	.58**	.20	.43**						
(e)	.44**	.40*	.44**	.35*					
(f)	.70**	.22	.50**	.45**	.46**				
(g)	.63**	–.01	.46**	.65**	.43**	.75**			
(h)	.42**	.26	.49**	.58**	.47**	.62**	.58**		
(i)	.46**	.15	.35*	.39*	.27	.13	.31*	.18	
Overall rating	.80**	.54**	.80**	.75**	.65**	.76**	.73**	.78**	.48**

*$p < .05$, **$p < .01$
(a) = Approach to Customers/Service, (b) = Attendance, (c) = Punctuality, (d) = Pace of work/Job motivation, (e) = Till procedure, (f) = Teamwork/Working relationships, (g) = Adaptability/Flexibility, (h) = Communication/Attitude, (i) = Standards of appearance. The average correlation between the three customer service dimensions is reported in the text.

Table 5. Correlations between the scales used in the customer service questionnaire and each of the supervisor scales.

	Scales of customer service			
Supervisor scales	D	E	S	Overall
Adaptability/flexibility	.15	.34	.20	.44*
Approach to customers/Service	.15	.24	.21	.43*
Standards of appearance	–.02	.09	.11	.36*
Attendance	.02	.11	.04	.40*
Communication/Attitude	–.09	.35	–.03	.39*
Pace of work/job motivation	.02	.42*	.10	.51**
Punctuality	.19	.34	.17	.59**
Teamwork/working relationships	.11	.24	.17	.38*
Till procedure	.32	.43*	.19	.62**
Overall supervisor rating	.13	.45*	.19	.74**

D = Dealing with people, E = Emotions and energy, S = Solitary style, Overall = Overall customer service assessment. Overall supervisor rating was the sum of the 9 job facet scales used to assess retail sales staff
*$p < .05$, **$p < .01$

and standard deviations of the supervisor ratings are also shown in Table 3. Also displayed is Cronbach's α, which is respectable for three of the four scales and adequate for the fourth scale ($\alpha = 0.68$).

Supervisor Ratings

After about 8 weeks on the job, 32 of the sales staff were rated by their immediate supervisors on nine qualities: Approach to Customers/Service, Attendance, Punctuality, pace of work/Job motivation, Till procedure, Teamwork/Working relationships, Adaptability/Flexibility, Communication/Attitude, and Standards of appearance. An overall supervisor rating was calculated as the sum of these nine scales (mean = 10.3, *SD* = 3.0, Number of items = 9). The remaining sales staff were not assessed by their supervisors as they had missed their assessment, left, or been transferred to another store. The supervisor ratings are the only method used to assess the quality of temporary sales staff in this organization. Low scorers are rejected by the organization should they apply for a full-time post.

Results

Table 4 shows that the supervisor ratings on each of the nine facets of the job were highly intercorrelated, which is suggestive of a high level of halo being present. The high Cronbach's α of the overall rating scale (= 0.85), which measures the internal consistency of the scale, confirms the view that the items were rated according to an overall impression.

Table 5 shows the correlations between D, E and S and the supervisor ratings. E was correlated with Pace of work/Job motivation ($r = 0.42$, $p < .05$) and with Till procedure ($r = 0.43$, $p < .05$). No other correlations between D, E, S, and the supervisor ratings were signifi-

cant. In contrast, the *overall scale* derived from the customer service questionnaire was correlated with *all* the supervisors' ratings on the facets and the overall rating ($r = 0.74$).

Discussion

In this study, three relatively orthogonal personality scales with generally respectable internal consistency were extracted from a well-designed selection questionnaire by means of principal components analysis. In general, these scales were *not* significantly correlated with supervisor ratings, supporting results of Furnham (1994) in which subscales from a customer service questionnaire were only shown to be modestly related to performance ratings but strongly related to personality scales, but contrasts with the general conclusions about the validity of this type of questionnaire drawn by Frei and McDaniel (1998).

At a higher-order level of analysis, an overall scale of customer service quality was, however, strongly correlated with all supervisor ratings on all job facets. Moreover, this overall scale had satisfactory internal reliability and was shown to be an effective summary of the three subscales that represent the content areas identified as being important in the job analysis. It is therefore reasonable to assume that this overall scale of customer service quality, derived from the selection questionnaire, is a reliable global measure of candidate potential. What could account for a high overall correlations with supervisor ratings which are not present at the subscale level? The best explanation stems from the demonstrable high levels of halo in the supervisor assessments. If supervisors also assess staff on a single global scale (as demonstrated by the high α coefficient of the overall supervisor rating scale and the very high correlations between the subscales), then it is likely that a single overall scale of retail sales potential would always be the optimal correlate as shown by Figure 1.

In Figure 1, the vectors of the subscales of the customer service questionnaire are at right angles to each other, which demonstrates a theoretical orthogonal relationship between the subscales. This is what is created by performing a principle components factor analysis and is a typical method for extracting subscales from a questionnaire since the orthogonality produces readily interpretable constructs. (In our study the subscales showed some intercorrelation, which resulted from unit weighting of only the useful items and the item analysis.) Since the vectors from these subscales of the customer service questionnaire are independent of each other, it is certain that many of the scales will not be related to supervisor ratings which contain high halo (i. e., which form just a single vector). However, as shown in Figure 1, the overall sum of the customer service subscales scales is more likely to be correlated to supervisors' global ratings. Figure 1 therefore provides a strong explanation of why there is only a low correlation between the supervisor's global impression and the orthogonal subscales of the customer service questionnaire (represented in the figure by the relatively large angle between the subscales and the supervisor assessments) and, in contrast, why there is a larger correlation between the two overall scales (shown by the small angle between the global customer service scale and the overall supervisor assessment).

A second explanation for the results appears quite weak in comparison: The greater choice of items that were available when designing the overall global scale of customer service may mean that items of greater validity were picked to be in that scale. In some cases this would be a strong possible explanation, but in this study it is an unlikely explanation since the items were assigned to the scale on the basis of classical item analysis which chooses items according to their reliability as opposed to their validity.

A third potential explanation comes from the personality literature. Ones and Viswesvaran (1996) discuss the bandwidth fidelity problem in personality measurement to which responses were made by Hogan and Roberts (1996) and Schneider, Hough, and Dunnette (1996). Essentially, Ones and Viswesvaran (1996) argue that *overall* personality constructs are more predictive of general performance measures than specific and narrow scales. While this approach has good potential to shed light on these results from a different perspective, our study shows that an *overall* personality measure is correlated with *individual* supervisor ratings as well as *overall* supervisor ratings. We therefore believe that our results are best explained using the "halo" explanation as opposed to the bandwidth argument, which is more of a description of the advantages of using high-bandwidth personality scales for personnel selection than an explanation of *why* high-bandwidth personality scales are useful. The halo explanation proposed in this study is in fact one good – and so far unexplored – explanation of Ones and Viswesvaran's observations.

The results of this study suggest that supervisor ratings with the halo rating error have utility as criteria in performance appraisal, because global assessments likely to contain the halo rating error appear to possess reasonable overall accuracy. This result should provide some element of comfort to those organizations that have informal appraisal systems using untrained appraisers. However, the results of this study should *not* be taken as general support for the use of supervisor ratings that contain the halo rating error. Halo is thought to result from

insufficient motivation, poor observation, poor quality rating instruments, lack of training, etc. – all of which suggests that the quality of supervisor ratings can be improved within such organizations, and ultimately this will be to their benefit because it would improve the accuracy of appraisals at the subscale level. When raters assess staff according to a global impression, the organization (a) lacks detailed knowledge of how its workforce is performing with the likely result of poor organizational planning and poor reward structures, (b) does not provide accurate feedback to staff with the likely result of high staff dissatisfaction, (c) and presumably therefore decreases sales. From this perspective, an organization would do well to ensure that the halo rating error is minimized. The negative effect that poor appraisal has on staff is certainly one possible explanation for the results of the job analysis in which it was found that poor attributional style was positively related to success. These negative correlations suggest that solitary, aggressive, pessimists who are high in emotional instability receive the better appraisals as well as those who are sociable, adaptable, assertive, active, satisfied, motivated and customer oriented.

Limitations

Halo often occurs when a rater rates according to a global impression, or in other words when the observed correlation between rating scales is higher than the true correlation. In this field study there were no known true correlations to subtract from the observed correlations. From a technical perspective, therefore, the observed correlations between the supervisor ratings (Table 4) were not halo, but in fact a mixture of true correlation and halo. A limitation of this study was, therefore, that an accurate measurement of the halo in excess of the true correlation was not available. However, a field study of this type will generally have this flaw in the design. After all, if true scores of performance were available in appraisal, then these would be used instead of ratings. Furthermore, the size of the intercorrelations shown in Table 4 is certainly much higher than the rating scales labels would suggest that they should be.

A second limitation to this study was that not all subjects who completed the customer service questionnaire were appraised by their supervisors. The inclusion of these staff in the correlation calculations may have led to very different results. The correlations between the subscales and the supervisor assessments may have been significant because restriction of range effects would have been reduced. Such a result seems possible; yet the results reported in this study represent the actual validity of staff who were still present at their appraisal time and thus were the only staff that the business was interested in appraising. The clarity of the results indicates that little would be gained by using a larger sample. A third limitation is that objective criteria (such as absenteeism and punctuality measures) were unavailable for this study because these measures were not made by the organization. Such measures would, however, have provided further knowledge about the supervisor assessments that were made in this study.

In summary, it has been argued that overall supervisor assessments used as criteria are *accurate* yet *flawed* global criteria of performance: They have a high correlation with a global selection measure but low correlations with more specific measures.

References

Balzer, W.K., & Sulsky, L.M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology*, *77*, 975–985.

Borman, W.C. (1977). Consistency of rating accuracy and rating errors in the judgement of human performance. *Organizational Behaviour and Human Performance*, *20*, 238–252.

Carmen, J. (1990). Consumer perceptions of service quality: An assessment of the SERVQUAL dimensions. *Journal of Retailing*, *66*, 33–35.

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

Costa, P., & McCrae, R. (1989). *The NEO-PI/NEO-FFI Manual supplement*. Odessa, FL: Psychological Assessment Resources.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, *66*, 127–148.

Fisicaro, S.A. (1988). A reexamination of the relation between halo error and accuracy. *Journal of Applied Psychology*, *73*, 239–244.

Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin*, *51*, 327–358.

Frei, R.L., & McDaniel, M.A. (1998). Validity of customer service measures in personnel selection: A review of criterion and construct evidence. *Human Performance, 11,* 1–27.

Furnham, A. (1994). The validity of the SHL customer Service Questionnaire (CSQ). *International Journal of Selection and Assessment*, *2*, 157–165.

Furnham, A., & Coveney, R. (1996). Personality and customer service. *Psychological Reports*, *79*, 675–681.

Hoffman, C.C., Nathan, B.R., & Holden, L.M. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self-versus supervisor ratings. *Personnel Psychology, 44,* 602–619.

Hogan, J., & Roberts, B.W. Issues and non-issues in the fidelity-bandwidth trade-off. *Journal of Organizational Behavior, 17,* 627–637.

Jackson, C.J. (1996). An individual differences approach to the

halo-accuracy paradox. *Personality and Individual Differences*, *21*, 947–957.

Kozlowski, S.W., & Kirsch, M.P. (1987). The systematic distortion hypothesis, halo and accuracy: An individual level analysis. *Journal of Applied Psychology*, *72*, 252–261.

Lance, C.E., LaPointe, J.A., & Stewart, A.M. (1994). A test of the context dependency of three causal models of halo rating error. *Journal of Applied Psychology*, *79*, 332–340.

Landy, F.J., & Farr, J.L. (1980). Performance rating. *Psychological Bulletin*, *87*, 72–107.

Murphy, K.R., & Balzer, W.K. (1986). Systematic distortions in memory-based behaviour ratings and performance evaluation: Consequences for rating accuracy. *Journal of Applied Psychology*, *71*, 39–44.

Murphy, K.R., & Cleveland, J.N. (1991). *Performance appraisal: An Organizational perspective*. Boston: Allyn and Bacon.

Nathan, B.R., & Alexander, R.A. (1988). A comparison of criteria for test validation: A meta-analytical investigation. *Personnel Psychology, 41,* 517–535.

Nathan, B.R., & Tippins, N. (1990). The consequences of halo "error" in performance ratings: A field study of the moderating effects of halo on test validation results. *Journal of Applied Psychology*, *75*, 290–296.

Ones, D.S., & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior, 17,* 609–626.

Parasuraman, A., Zeithaml, V., & Berry, L. (1985). A conceptual model of service quality and its implications for future research. *Journal of marketing*, *49*, 41–50.

Parasuraman, A., Zeithaml, V., & Berry, L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, *67*, 420–450.

Schneider, R.J., Hough, L.M., & Dunnette, M.D. (1996). Broadsided by broad traits: How to sink science in five dimensions or less. *Journal of Organizational Behavior, 17,* 639–655.

Chris Jackson
School of Psychology
University of Queensland
Brisbane QLD 4072
Australia
E-mail chrisj@psy.uq.edu.au