Pergamon

Person. individ. Diff. Vol. 21, No. 6, pp. 947–957, 1996

0191-8869/96 $15.00+0.00

S0191-8869(96)00144-4

AN INDIVIDUAL DIFFERENCES APPROACH TO THE HALO-ACCURACY PARADOX

Chris Jackson

Department of Psychology, London Guildhall University, Old Castle Street, London E1 7NT, England

(Received 26 January 1996)

Summary—In performance appraisal, the halo-accuracy paradox describes the surprising result that rater accuracy can be *positively* correlated with the halo rating error. Fisicaro (1988) provided an explanation for this unlikely relationship by proposing an inverse V function as the relationship between accuracy and invalid halo in which maximum accuracy is located at zero invalid halo. This paper develops the model by proposing that maximum accuracy does not have to be at zero invalid (Hypothesis 1). As the cognitive difficulty of a rating task increases, a negative monotonic relationship between maximum achievable accuracy and associated value of absolute invalid halo is specified (Hypothesis 2). The hypotheses were tested in two different experimental situations. Results from both studies supported Hypothesis 1 but, whilst a distinct pattern between accuracy and absolute invalid halo was noted, only a weak version of Hypothesis 2 could be supported. The evidence from this paper demonstrates that the halo-accuracy paradox is not an artefact as some recent reviewers have proposed (Balzer & Sulsky, 1992; Murphy & Balzer, 1989; Murphy & Cleveland, 1991). Copyright © 1996 Elsevier Science Ltd.

INTRODUCTION

In performance appraisal, workers are often assessed on numerous rating scales that reflect the different personal qualities that are important to the job. Halo has traditionally been seen as a type of *rater error* that 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* (see Borman, 1977; Fisicaro, 1988; Lance, LaPointe & Stewart, 1994). Such a definition has been extended by theorising (Cooper, 1981; Feldman, 1981; Landy & Farr, 1980) that the halo results from raters rating according to their own models of the world (i.e. ratings are systematically distorted as a result of implicit theories, person schemata or prototypes). As such, observed correlations between rating scales need not always be higher than the true correlations and can sometimes be lower (Murphy & Reynolds, 1988). The positive or negative difference between observed and true correlation is known as *invalid halo*.

The aim of using rating scales in performance appraisal is to achieve accurate assessments that reflect the true ability of the worker. Occasionally, a positive relationship between accuracy and invalid halo has been observed. This is called the *halo-accuracy paradox* (Cooper, 1981; Murphy & Cleveland, 1991), because it is surprising that an increase in invalid halo rating error is sometimes associated with an increase in rating accuracy (Jackson, 1989a; Kozlowski & Kirsch, 1987; Murphy & Balzer, 1986; Nathan & Tippins, 1990).

Fisicaro (1988) attempted to resolve the paradox 'within' tasks by hypothesising an inverse V relationship between accuracy and invalid halo, with maximum accuracy at the point of zero invalid halo. This translates into a negative monotonic relationship between accuracy and absolute invalid halo (Fig. 1). With this model, positive and negative correlations between accuracy and invalid halo arise from differing proportions of raters with positive and negative invalid halo. However, when the *absolute* value of invalid halo is used, Fisicaro reported only negative correlations. This was thought by Fisicaro to resolve adequately the paradox.

However Fisicaro's model does not explain *positive* correlations between accuracy and *absolute* invalid halo within tasks and makes no attempt to explain between task effects. Despite the evident strengths of Fisicaro's model, these limitations may explain why the relationship between these variables still confuses researchers (Balzer & Sulsky, 1992). This confusion has led some recent

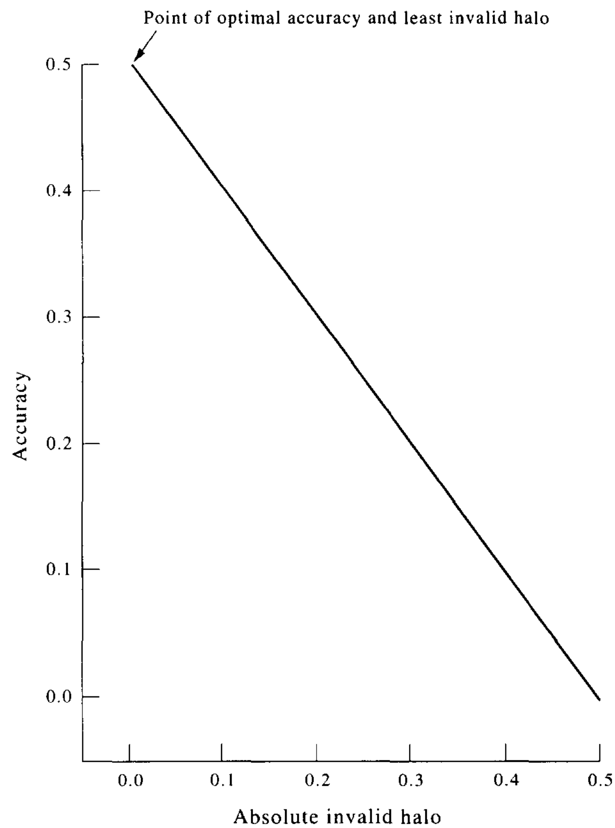

Fig. 1. Fisicaro's model of the relation between accuracy and absolute invalid halo.

reviewers to conclude that the halo-accuracy paradox is an artefact resulting from poor measurement techniques (especially when the problems of measuring true scores are considered) and therefore that Fisicaro's model is of little practical relevance anyway (Murphy & Balzer, 1989; Murphy & Cleveland, 1991).

This paper aims to demonstrate how Fisicaro's model can be extended to explain: (1) both positive and negative correlations between raters' accuracy and absolute invalid halo *within* tasks; (2) negative relationships between maximum accuracy and associated value of absolute invalid halo *between* tasks; and (3) why the reviewers' conclusions may therefore be premature.

Fisicaro's model proposed that maximum accuracy is achieved at zero invalid halo, implying that a rater rates most successfully when using *all* available information in a *perfect* manner. In a review of cognitive models of the rating process, in particular of Feldman (1981), Lee (1985) asserted that raters "have limited acquisition, storage, retrieval and integration capacities" (p. 322). This cognitive perspective indicates a limit well below the theoretical level that Fisicaro's model demands, because the optimal processing capacity of even the most accurate rater may not *necessarily* be at zero invalid halo. If maximum accuracy is not necessarily at zero invalid halo (Fig. 2), then one of the limitations of Fisicaro's model will have been identified. Hypothesis 1 states that *within* a rating task, individual raters who achieve maximum accuracy will not necessarily have an associated value of zero invalid halo.

As tasks become more difficult, it is likely that there will be a reduction in accuracy and increase in rater error. Hypothesis 2 states that maximum achievable accuracy *between* tasks will decrease and absolute level of invalid halo associated with it will increase, as the cognitive difficulty of the task increases. Direction of effects was specified *a priori*, because "theoretically guided hypotheses about information processing", ... are essential, or else it ... "can lead to inappropriate interpretations and should therefore be avoided" (Balzer & Sulsky, 1992, p. 983).

Hypotheses 2 is illustrated by combining Fig. 1 with Fig. 2. Maximum accuracy of the task represented by Fig. 1 is higher than that of Fig. 2 and is associated with a lower value of absolute

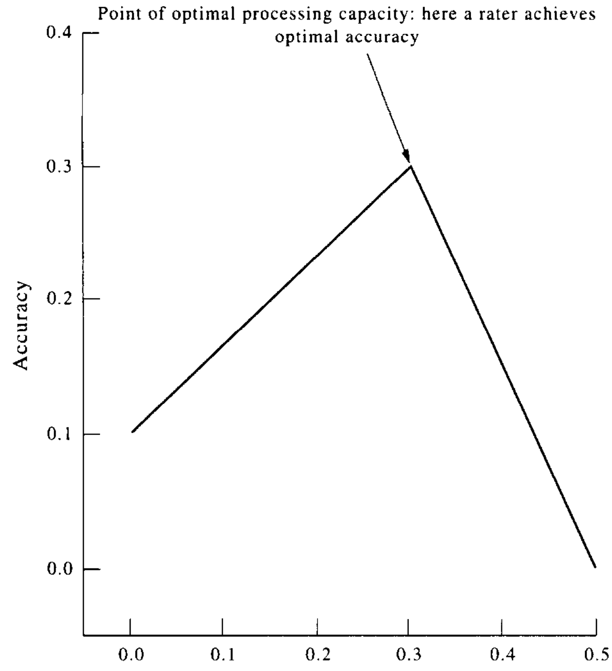

Fig. 2. Extension to Fisicaro's model in which the maximum is not necessarily at zero invalid halo.

invalid halo. In other words, the task represented in Fig. 1 is *less* cognitively demanding than the task represented in Fig. 2. These graphs show that Fisicaro's model is likely to be only appropriate for easy rating tasks in which high maximum accuracy and low associated absolute invalid halo is found.

Experimental design

Study 1 was an artificial, laboratory investigation in which the weight, height and age of individuals in photographs were rated. Although these subjective assessments are simpler constructs than the vaguer notions that are generally rated in real-life performance appraisal (e.g. managerial potential, aircraft handling properties or gymnastic ability), the experimental design provided *direct access* to true scores.

The design of Study 2 was more realistic. Videotapes were made of interviews that are used to select candidates for training within a large uniformed organisation (Dexter, 1984; Jackson, 1989b; Jackson, 1995). Estimated true scores were derived by averaging expert assessments of these videotapes; a *validated* method *widely used* in the performance appraisal literature (Borman, 1977; Smither, Barry & Reilly, 1989).

Halo error and accuracy can be measured in various ways (Fisicaro, 1988; Saal, Downey & Lahey, 1980). In Study 1, with multiple *S*s and multiple stimuli, it was possible to measure absolute invalid halo and accuracy as a correlation (Kozlowski & Kirsch, 1989; Lance, Fisicaro & LaPointe, 1990). However, in Study 2, limited numbers of *S*s and videotapes meant that correlational measures could not be applied. Instead, standard deviation measures were used to assess invalid halo (similar to that used by Nathan & Tippins, 1990) and a distance measure was used to assess accuracy. This is consistent with conceptual and psychometric definitions of accuracy (Zalesny & Highhouse, 1992).

PROCEDURE OF STUDY 1

Subjects

The sample comprised 47 males and 53 females. All were students between 19 and 27 years of age.

Table 1. Experimental design of Study 1

	Task A rating only	Task B rating with secondary task
1st presentation	Task A (1)	Task B (1)
2nd presentation	Task A (2)	Task B (2)

1st presentation was with no practice effect, whereas 2nd presentation involved a practice effect.

Method

A selection of colour photographs of 42 males and females of known height, weight and age were used in the study. The photographs showed people in natural poses in different situations and with different backgrounds. The aim was to create a relatively mixed and random set of photographs. *S*s were not acquainted with the individuals in the photographs.

*S*s were randomly split into two groups of equal size. The first group viewed a total of 21 photographs in random order, each for 5 sec. After viewing each photograph, *S*s rated the height, age and weight of the individual in the photograph [results classified as Task A(1)]. Next, the *S*s viewed the remaining 21 photographs and made the same assessments but this time also counted backwards during the 5 sec of observation [results classified as Task B(2)]. *S*s in the second group performed the same two rating tasks but in reverse order; Task B followed by Task A [classified as Task B(1) and Task A(2) respectively]. The 2×2 classification of the four rating tasks is shown in Table 1.

It was expected that the presence of a secondary task would reduce accuracy and increase rater error compared to making just a rating. Also, it was expected that practice should reduce demands upon cognitive processing, thereby increasing accuracy and reducing rater error. Making just a rating with a practice effect was therefore specified *a priori* as the cognitively 'easy' condition in this study, whereas presence of a secondary task and no practice was specified as the 'hard' condition.

RESULTS OF STUDY 1

For 'each individual rater' in each of the four rating tasks, measures of average absolute invalid halo and average accuracy were calculated using the r to z transform as necessary. The calculation of average absolute invalid halo for each individual rater was:

$$\mathrm{INV \cdot H_{AV \cdot ABS} = AVERAGE[ABSOLUTE(OBS \cdot H_{pp} - TRU \cdot H_{pp})]}$$

for $P = 1 \ldots 3$ (age, height and weight) and $\mathrm{INV \cdot H_{AV \cdot ABS}}$ = The absolute invalid halo correlation calculated as an average for each rater; $\mathrm{OBS \cdot H_{pp}}$ = The observed correlation between the ratings of $p_{1,2}$, $p_{1,3}$ and $p_{2,3}$; $\mathrm{TRU \cdot H_{pp}}$ = The true correlation between $p_{1,2}$, $p_{1,3}$ and $p_{2,3}$. Accuracy of each individual rater was calculated as the average of the correlations between the true scores and respective ratings.

A one way ANOVA of absolute invalid halo (Table 2) indicated significant differences between the four experimental conditions ($P = 0.00$). Examination of the confidence intervals of the means indicated no differences between Task A(1) and Task A(2). Therefore differences between only three rating tasks were examined [Task A(1,2), Task B(1) and Task B(2)]. A scatter-plot of the relationship between each individual rater's accuracy and absolute invalid halo was also produced for Task A(1) and Task A(2). This confirmed that there were no apparent differences between these two tasks

Table 2. ANOVA to determine if there was a difference between average absolute invalid halo for each of the rating tasks of Study 1

Source	d.f.	SS	*F*	*P*
Task	3	0.81	9.37	0.000
Error	196	5.68		
Total	199	6.49		

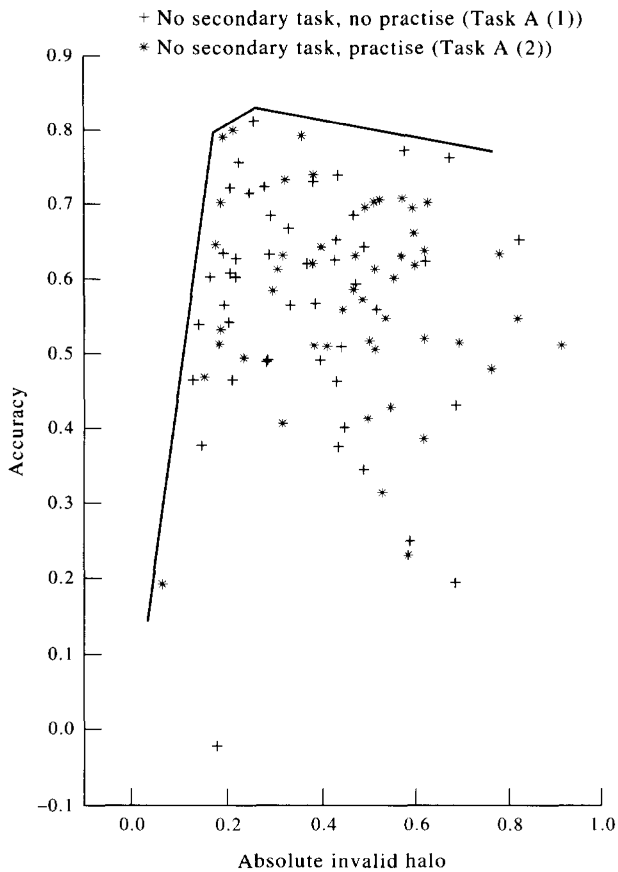

Fig. 3. Scatter-plot of accuracy against absolute invalid halo for Study 1 using data from Task (A1) and Task (A2).

(Fig. 3). Note, there are no datapoints in Fig. 3 indicating highest accuracy was achieved by any rater at zero invalid halo. It can also be seen that data-points appear not to lie on a curve, but to be located *on* or *under* a curve.

Summary statistics (Table 3) showed that maximum accuracy was not located at zero invalid halo for any of the three tasks. Each task had a different maximum accuracy and as this increased in size, the value of absolute invalid halo associated with it decreased.

Other statistics were also summarised in Table 3. The correlation between accuracy and absolute invalid halo tended to become increasingly negative as maximum accuracy increased and its associated value of invalid halo decreased. The range of absolute invalid halo tended to become smaller as maximum achievable accuracy increased. Also noted is the large number of ratings that were required to measure accuracy and absolute invalid halo as correlations.

Table 3. Summary statistics of Study 1

Task	K	Hmax	Rmax	*r*	No. of outliers	Range of halo values	*M*
Task A (1,2)	100	0.24	0.81	−0.034	0	0.1–0.9	63
Task B (1)	50	0.17	0.94	0.019	1	0.1–0.8	63
Task B (2)	50	0.11	1.00	−0.352	0	0.1–0.7	63

Task A (1,2) involved rating age, weight and height from photographs after 5 sec of observation. Order of presentation was not a significant factor and so Task A was not split into (1) and (2).

Task B(1) and B(2) involved rating age, weight and height from photographs after 5 sec of observation whilst counting backwards at the same time. The order of presentation is shown by (1) (no practice) and (2) (practice).

K, Number of raters; Hmax, The absolute invalid halo value associated with Rmax; Rmax, The maximum accuracy achieved by subjects; r, The correlation between accuracy and absolute invalid halo (all data included); M, Total number of ratings made per rater.

Table 4. Dimensions of performance that were important and observable in the interview

Appearance and bearing
Manner and impact
Powers of expression
Breadth and depth of sports activities
Breadth and depth of other activities
Educational background and academic potential
General knowledge and awareness
Maturity of character
Motivation

PROCEDURE OF STUDY 2

Study 2 consisted of two stages. First, the development of the true scores related to eight interviewees on videotape; and, second, the rating procedure.

Stage 1

The first stage is described elsewhere (Jackson, 1989b; Jackson, 1995) but the essential details are recorded here. A job analysis conducted by Subject Matter Experts (SMEs) identified nine dimensions of performance that were important to success and that were also observable in the videotaped interviews (Table 4). Each dimension was rated in terms of how important and how trainable it was. By dividing importance ratings by trainable ratings and then averaging across SMEs, an estimated true score was derived. The correlation between these scores was used as an estimate of valid halo. None of the correlations was significant. The SMEs also provided one overall true score rating for each interviewee on a 1–15 point scale of potential for success at the next stage of training.

Stage 2

Subjects. The sample comprised 32 trained and job knowledgeable interviewers. None of these assessors were involved in the first stage.

Method. A Graeco–Latin square experimental design was used so that each group of interviewers watched four interviewees on videotape and performed only one rating task per viewing of an interviewee. The experimental design controlled for possible order effects derived from interviewees or rating tasks.

The 2×2 design of the four rating tasks is shown in Table 5. The rating tasks were as follows:

(1) Make notes on each dimension, rate each dimension on a 0–7 scale, summarise the notes and then award a 0–7 Overall Assessment Rating (OAR). This task was classified as Task Notes (8).
(2) Make notes on each dimension, rate each dimension on a 0–1 scale, summarise the notes and then award a 0–7 OAR, classified as Task Notes (2).
(3) Rate each dimension on a 0–7 scale and then award a 0–7 OAR, classified as Task no Notes (8).
(4) Rate each dimension on a 0–1 scale and then award a 0–7 OAR, classified as Task no Notes (2).

It was expected that the condition of No Notes and Two point rating scales would show lowest accuracy and highest invalid halo (i.e. the cognitively 'hard' condition) when compared with the condition of Notes and Ten point rating scales (i.e. the 'easy' condition). Notes could be expected to reduce cognitive demands because they summarise complex observations. Ten-point-rating scales should reduce cognitive demands because they are similar to the kind of assessments that the raters were used to making.

Table 5. Experimental design of Study 2

	Notes	No notes
0–7 Rating	Task Notes (8)	Task no Notes (8)
0–1 Rating	Task Notes (2)	Task no Notes (2)

Table 6. Summary statistics of Study 2

Rating task	K	Hmax	Rmax	*r*	No. of outliers	Range of halo values	L	*M*
Task Notes (8)	32	1.06	2.25	−0.57	1	0.8–1.8	2	36
Task Notes (2)	32	1.20	1.95	−0.022	2	0.9–1.6	11	36
Task no Notes (8)	32	1.08	1.95	0.158	0	0.6–3.6	2	36
Task no Notes (2)	32	1.31	1.35	0.370	2	0.9–1.7	9	36

Task Notes (8) = Notes/eight point rating scale.
Task Notes (2) = Notes/two point rating scale.
Task no Notes (8) = No notes/eight point rating scale.
Task no Notes (2) = No notes/two point rating scale.
Hmax, Level of invalid halo corresponding to maximum accuracy; Rmax, Maximum accuracy.
r, Linear correlation between accuracy and invalid halo; K, Number of raters; L, Number of disregarded points (see text); M, Number of ratings made per rater.

RESULTS OF STUDY 2

To ensure there were no statistical differences between binary and polychotomous scales (Cohen, 1983), polychotomous ratings were recoded as dichotomised data. All data were then standardised to remove mean differences (Pulakos, Schmitt & Ostroff, 1986). Subsequently, invalid halo was computed as the inverse, observed, standard deviation between dimensions (Fisicaro, 1988; Saal *et al.*, 1980). In this study, valid halo was not an important factor since Stage (a) had provided evidence that its effect was not significant.

Accuracy was calculated as the absolute inverse standardised difference between the OAR and the overall estimated true score. Hence a rater was recorded as showing greater accuracy if the positive or negative difference between the awarded OAR and the estimated true score was small compared to when it was large. This measure of accuracy is similar to the approach of Zalesny and Highhouse (1992).

In general, maximum accuracy of each task was not at zero invalid halo and as maximum accuracy increased, the following occurred (Table 6): (1) the level of invalid halo associated with maximum accuracy decreased; (2) the correlation between accuracy and invalid halo became more negative; and (3) the range of invalid halo did not appear to be related to maximum accuracy.

Scatter-plots of accuracy against invalid halo showed that data appeared to be randomly distributed on or under curves that were drawn by visual inspection to enclose the data. The number of outliers and the number of disregarded datapoints are recorded in Table 6. Disregarded datapoints occurred when raters made the same ratings on all dimensions; a problem commonly found when using dichotomised data.

DISCUSSION

Rater accuracy and invalid halo

Both studies provided strong evidence that maximum accuracy *within* a task was not necessarily at zero invalid halo. Thus, for example, the distinct curve drawn in Fig. 3 does not have a maximum at zero invalid halo. Scatter-plots of the two other tasks in Study 1 and the four rating tasks in Study 2 all showed distinct absence of points above a curve and, where appropriate, a distinct absence of points illustrating highest accuracy at zero invalid halo. Fisicaro's model therefore appears to be a special case of a more general model in which maximum accuracy is not necessarily associated with zero invalid halo. Hypothesis 1 was therefore supported.

Between rating tasks, there was a strong tendency for a negative monotonic relationship between maximum accuracy and absolute invalid halo. This evidence only partially supported Hypothesis 2, because the relationship between these variables was not always in the expected direction. Results of Study 1 provided evidence that inclusion of a secondary task and practice could improve rating quality. Practice is likely to make a rating task *less* demanding, but the effect of a secondary task is counter-intuitive since performing two tasks could be expected to be *more* demanding than performing one. Further research may well explain this effect, although it may be due to the change of the rating process from conscious to unconscious processing (Feldman, 1981) when the rater's attention is not directly focused on performing just the rating task.

Results of Study 2 indicated that the use of notes and eight-point-rating scales was generally the most accurate and subject to least invalid halo error (i.e. was the least cognitively demanding) of the four rating tasks. Using no notes and two point rating scales was the least accurate and contained most error (i.e. was the most cognitively demanding). Notes with two point scales and no notes with eight point scales were in between these extremes. Thus the direction of the cognitive effects of Study 2 were as predicted by Hypothesis 2 and there is evidence that notes and polychotomous rating scales can improve rating quality.

Since maximum achievable accuracy of a task appears to be related to its associated level of absolute invalid halo, the results provide evidence in support of Murphy, Jako and Anhalt's (1993) statement, "whether halo errors should be suppressed or avoided depends largely on the context in which ratings are elicited or ratings are obtained" (p. 223). Practical implications of these studies are: (1) it should be beneficial to encourage halo when the rating task is cognitively demanding; and (2) rating tasks should be kept as simple as possible.

TOWARDS A GENERAL MODEL

The following points can also be concluded from the two studies: (1) Even within a task, there are wide individual differences between raters with regard to their levels of invalid halo; (2) Plots of raters' accuracy against absolute invalid halo for each rating task illustrated that data tended to be 'on' or 'under' a curve; and (3) As the level of absolute invalid halo associated with maximum accuracy increased, the correlation between accuracy and absolute invalid halo changed from negative value, through zero, to positive value.

Individual differences between raters

This paper has argued in favour of a single peaked function relating accuracy to absolute invalid halo with a maximum equivalent to rater's optimal processing capacity. Not all raters will use their optimal information processing capacity to make a rating. If processing capacity is not reached (for example, an unmotivated rater rating according to an overall impression), there would be relatively high levels of absolute invalid halo and reduced accuracy, in comparison with the optimal, because of an increased level of cognitive distortion within the rating process. This is the likely reason why raters with a higher than optimal level of absolute invalid halo tended to have reduced accuracy.

When processing capacity is exceeded, there is likely to be a lower level of absolute invalid halo than the optimal, because the rater has attempted to utilise more processing capacity than is available. If this is the case, then it is likely that the rater will become overloaded and thus 'less' accurate than the optimal. It is unclear how information overload may decrease rater performance but, in different contexts, overload can result in dysfunctional performance strategies (Huber, 1985) or confusion (Sales, 1970). Information overload may also decrease accuracy by increased reliance on non-linear decision making.

Data lie on or under a curve

Evidence shows that data-points lie on or under a boundary curve. Leniency and central tendency rating errors are likely to reduce accuracy whilst not overly effecting absolute invalid halo (Jackson, 1989a, Murphy & Cleveland, 1991; Saal *et al.*, 1980). The presence of other types of rater error explain why datapoints are found below a curve.

Accounting for the halo-accuracy paradox and proposing a general model

It is now possible to propose a general model that will explain the results of these two studies. Consider three examples of the proposed model (illustrated in Fig. 4). Hypothesis 1 is illustrated by noting that maximum accuracy of each of the three example tasks is not at zero invalid halo. Hypothesis 2 is represented by Line A. Now examine each of the examples more closely. First,

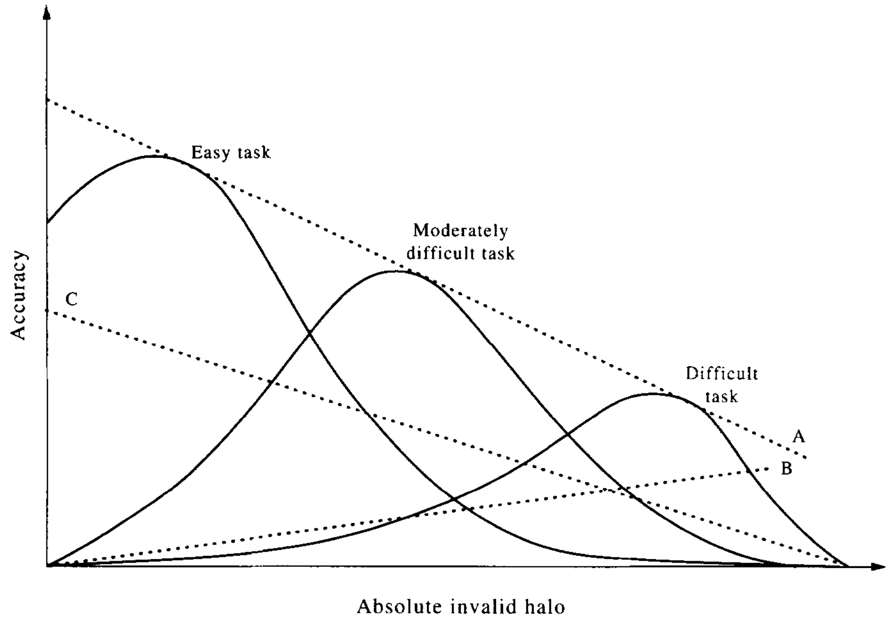

Fig. 4. Proposed relationship between accuracy, absolute invalid halo and task difficulty. Each curve represents the boundary line under which points will be scattered. The maximum of each curve represents rater capacity. The curves result from the reduction in accuracy when capacity is overloaded and the reduction in accuracy as a result of underutilising cognitive resources. A, Represents the negative overall relationship between maximum achievable accuracy and associated absolute invalid halo over tasks with different cognitive demands. B, Represents the positive correlation between accuracy and absolute invalid halo for cognitively demanding tasks. C, Represents the negative correlation between accuracy and invalid halo for less cognitively demanding tasks.

consider the 'easy' rating task in which low absolute invalid halo equates to optimal accuracy. Except for the cases of very low invalid halo, there is little chance of information overload. However, some raters still show relatively high absolute invalid halo, and thus low accuracy. The location of the maximum explains the observed negative correlation between absolute invalid halo and accuracy for less demanding tasks.

Now consider a cognitively demanding rating task. In this case, the optimal rater could be expected to show relatively high absolute invalid halo because information overload has been avoided. Some raters would have lower absolute invalid halo than the optimal and some would have higher than the optimal, but the location of the maximum would cause a *positive* correlation between absolute invalid halo and accuracy, as observed in the results from the two studies. Figure 4 displays the changes in the correlation between accuracy and absolute invalid halo. Line B illustrates the positive correlation between accuracy and invalid halo for cognitively demanding tasks and Line C illustrates the negative correlation between accuracy and invalid halo for less demanding tasks. Following this model, a positive correlation between accuracy and absolute invalid halo is no longer a paradox.

This paper has proposed an individual differences approach to resolving the halo-accuracy paradox. Its aim was to improve present knowledge of the relationship between accuracy and absolute invalid halo by extending Fisicaro's model and by suggesting that the halo-accuracy paradox is not necessarily an artefact as recent researchers have proposed (Murphy & Cleveland, 1991). If invalid halo can be beneficial, perhaps it is better interpreted in terms of individual differences between raters who adopt various 'strategies' or 'styles' to accomplish their tasks rather than just committing rater errors. Choice of word, error or strategy, has important practical ramifications in terms of the approach that a trainer of raters might use. An error infers the need for correction to zero level, whereas a strategy might indicate that calibration to some other level is required.

LIMITATIONS OF THE STUDIES

In Study 1, the need to have direct access to true scores was of paramount importance. This meant that the rating tasks were rather artificial, and so these ratings may not be comparable to ratings made in the field. Additionally, using trained assessors limited the number of *S*s in Study 2 and thus the power of the design. It is also believed that valid halo in Study 2 could have been more accurately estimated, although for this study, it was not possible. It should also be noted that the use of experts in Study 2 to estimate true scores is limited because the true scores are subjective assessments that are unlikely to have a correlation of 1.0 with success in passing training. Nevertheless it is likely that these optimal ratings will equate to true scores because rater error is likely to have been removed as a result of averaging between experts and the technique has been shown to have validity (Smither *et al.*, 1989). Finally, note that in these studies, rating accuracy was not partitioned into its components (as proposed by Cronbach, 1955) because the aim was to examine the 'overall' relationship between invalid halo and accuracy.

When examining plots of the relationship between accuracy and invalid halo for each rating task, the curves can only be drawn approximately as there is no statistical method to fit a boundary line to enclose a set of data. As such, drawing a curve to enclose data by visual inspection will always be problematical, because it is not clear if a data-point lies on the curve or under it.

Acknowledgements—Thanks to Sandy Kalia for helping to prepare the text and David Diamond for collection of the data in the first study.

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.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–253.

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

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

Dexter, R. A. (1984). Officer selection in the eighties. *Air Clues*, September, 348–350.

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

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

Huber, V. L. (1985). Effects of task difficulty, goal setting, and strategy on performance of a heuristic task. *Journal of Applied Psychology*, *70*, 492–504.

Jackson, C. J. (1989a). *Relations between rater bias, accuracy and perceived task difficulty*. Unpublished Ph.D. Thesis, Coventry University, Coventry.

Jackson, C. J. (1989b). *Improvement to the subjective rating methods used to select Officers for the RAF*. Science 3 (Air), Ministry of Defence, Lacon House, Theobalds Road, London.

Jackson, C. J. (1995). Assessing important and observable personal qualities in the general selection interview. *European Journal of Psychological Assessment*, *11*, 75–80.

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., Fisicaro, S. A. & LaPointe, J. A. (1990). An examination of negative halo error in ratings. *Educational and Psychological Measurement*, *50*, 545–554.

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.

Lee, C. (1985). Increasing performance appraisal effectiveness: Matching task types, appraisal process, and rater training. *Academy of Management Review*, *10*, 322–331.

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. & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*, 619–624.

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

Murphy, K. R., Jako, R. A. & Anhalt, R. L. (1993). Nature and consequences of halo: A critical analysis. *Journal of Applied Psychology*, *78*, 218–225.

Murphy, K. R. & Reynolds, D. H. (1988). Does true halo affect observed halo? *Journal of Applied Psychology*, *73*, 235–238.

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.

Pulakos, E. D., Schmitt, N. & Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within rates to measure halo. *Journal of Applied Psychology*, *71*, 29–32.

Saal, F. E., Downey, R. G. & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428.

Sales, S. M. (1970). Some effects of role overload and role underload. *Organizational Behaviour and Human Performance*, *5*, 592–608.

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

Zalesny, M. D. & Highhouse, S. (1992). Accuracy in performance evaluations. *Organizational Behaviour and Human Decision Processes*, *51*, 22–30.