Journal of Industrial Psychology, 1993, 19(1), 11-14
Tydskrif vir Bedryfsielkunde, 1993, 19(1), 11-14

TOWARDS THE STANDARDIZATION OF AN INSTRUMENT FOR THE EVALUATION OF TRAINING

C J H BLIGNAUT

Department of Human Resource Management
Rand Afrikaans University

ABSTRACT

A review of the literature indicates that most evaluations of the effectiveness of training are performed at the reaction level of Kirkpatrick's model. Although the model is generally accepted, evaluation at the reaction level of the model is not. The present study is aimed at resolving this paradox. In particular it is argued that the incongruous results originate from inadequate instruments. The instrument developed in this study, is based on the principles of perception and learning and was found to be psychometrically sound for application in research and practical settings.

OPSOMMING

'n Oorsig van die literatuur toon dat die evaluering van opleiding meestal op die reaksievlak van Kirkpatrick se model gedoen word. Alhoewel die model oor die algemeen aanvaar word, is dit nie die geval met evaluasies op die reaksievlak van die model nie. Die studie is daarop gerig om die paradoks uit te skakel. In besonder is daar geredeneer dat genoemde onversoenbaarheid aan ontoereikende meetinstrumente toegeskryf kan word. Die instrument wat in die studie ontwikkel is, is op die beginsels van waarneming en leer gebaseer en blyk psigometries aanneemlik te wees vir gebruik in navorsings- en praktiese omstandighede.

In the training and development literature frequent reference is made to the importance and necessity of evaluating the effectiveness of training. Some of the reasons offered relate to organisational issues, such as whether the return on the training investment is worthwhile, whether the training effort contributes to organisational improvements and objectives, and whether training programmes result in the acquisition of those skills which the organisation will need over the short term (Bell & Kerr, 1987). Other reasons seem to relate to training itself: the design of a measuring instrument forces attention to the relevance and clarity of training objectives; evaluation conveys the message to programme participants that interest is shown in results through learning and behaviour changes (Quinn & Karp, 1986). Ultimately though, it seems as if "trainers must demonstrate training's value in more substantive ways if it is to *survive and* gain its rightful place among investment alternatives" (Carnevale & Schultz, 1990; italics added).

The theoretical discourse in the area of training evaluation tends to focus on the presentation and explanation of models of evaluation. In this regard three widely acclaimed approaches or models can be identified in the literature. The first relates to traditional research methodology where issues such as control groups and pre/post-test designs are discussed (Goldstein, 1974). The second relates to the description of evaluation in terms of general systems theory. From this perspective evaluation is seen as the formalisation of processes through which feedback is made available to management and trainers on the basis of which corrective action may be taken. The third involves the classification of levels of evaluation as proposed by Kirkpatrick (1979). From the numerous references made to Kirkpatrick, it can be concluded that his model is entrenched (whether directly or indirectly) in the area of training, so much so that it is seen as a classical contribution (Parker, 1986) and that it has been accepted as a framework in industrial/organisational psychology (Cascio, 1987). In short, Kirkpatrick proposed that the evaluation of training take place at one or more of the following levels: *reaction* (how did participants like the training programme?), *learning* (what knowledge did participants gain from the programme?), *behaviour* (what job behaviour changes resulted from the programme?) and *results* (what effects did the programme have on the organisation?

The author wishes to acknowledge the help of Prof JM Schepers in finalising the study.
Requests for reprints should be made to the author at P.O. Box 524, Aucklandpark 2006.

Whereas the reasons for and theory of evaluation are convincing and sound, there is ample evidence in the literature that, in practice, the actual evaluation of training lags behind. For example, Carnevale and Schultz (1990) suggest that fewer than half of the training programmes in the USA are evaluated. Empirical evidence in this regard stems from the studies of Clement and Walker (1979), Smeltzer (1979) and Carlisle (1984). Parker (1986) again clearly demonstrated that the majority of companies tend to focus their training evaluation on the reaction level rather than on the other levels of the Kirkpatrick model.

Reaction evaluation, however, has received mixed approval in the literature. Morris (1984), for example, enumerated the limitations of reaction evaluation, whilst Birnbrauer (1987) is of the opinion that reaction to training cannot be validated and that there is little reason to assume, as was confirmed by Alliger and Janak (1989), that reaction would correlate with actual learning. On the other hand Whitelaw (1972) considers reaction-level evaluation as immediate feedback which can be quite useful to trainers. Newstrom (1978) notes that the assumption is often made that: "If trainees react favourably, they'll probably learn more" (p. 22), thereby implying that there is a positive relationship between the first and second levels of Kirkpatrick's hierarchy.

The diversity of opinions regarding the usefulness of learner reaction and the general scepticism with which it is viewed, probably stems from the fact that evaluation models tend to be proposed, explained and conceptually assessed rather than empirically tested. Specifically, it is suggested that the dimensions by which reactions tend to be evaluated, are inappropriate and that the construction of reaction-level measures are not psychometrically sound. From the literature, for example, the impression is gained that trainees are required to respond to items which relate to the personality and skills of the trainer, to the available training facilities, programme content and to administrative arrangements. Furthermore, evaluation is performed in such a way that an indication of the likes and dislikes of trainees is obtained. It thus came to pass that some authors, for example Bell and Kerr (1987), began to apply the term "happiness index" to evaluation at the reaction level.

There seems to be a paradox embedded in the above: on the one hand Kirkpatrick's model is generally accepted and hailed as a landmark in the area of training evaluation. Yet, on the other hand, the very first level of evaluation in the model

(trainee reaction) appears to be rejected or at best viewed with scepticism. Another incongruency is also apparent: although first-level evaluation is severely criticized on theoretical grounds, it remains the basis for most evaluation research carried out by trainers. These discrepancies are the focus of this study. In particular it is suggested that the discrepancies do not necessarily result from Kirkpatrick's model *per se*, but that they probably arise from the research side, that is, from the nature of the items characteristic of reaction measures as well as from the apparent absence of psychometric considerations during the design of such measures. The aim of this study then, is to construct a psychometrically acceptable instrument at the reaction level of Kirkpatrick's model and, in this regard, to make use of items of which the content has been shown to relate to learning.

METHOD

Rationale underlying the proposed instrument

Instrument design requires considerable clarity regarding the purpose of the instrument and the construct to be measured. In the present study it was argued that the "reaction of trainees" implies that their responses to a training session or programme should be measured in terms of the "learning opportunities" which they perceived were present in the session or programme. Learning opportunities, for the purposes of this study, were specified in terms of the principles established in the areas of perception and learning. Regarding the former area it was argued that training essentially implies the transfer of information to trainees. As such the speed, accuracy and ease with which trainees receive information in training situations seems to be of critical importance. This is evident not only from systems theory when applied to information processing in human beings, but also from psychology where perception is regarded as the primary and only means through which the world is known. To elaborate somewhat: information processing is set in motion with the input or perception of information. It is asserted that the learning process is also set in motion through the perception of information by trainees. Furthermore, in the study of "man-machine systems" it has been demonstrated clearly (Wickens, 1992) that the way in which information is presented, profoundly affects the speed and accuracy of the operator's decisions and actions. It is posited that the presentation of information in training situations will have a similar effect on the learning process. In literature on the psychology of perception, it is clear that the effects referred to above, result from factors associated with "attention" and the "sensory" and "perceptual" mechanisms of the individual. The dominant role of factors such as stimulus intensity, contrast, similarity, redundancy and amount of information have been demonstrated to the extent that they are in general referred to as the principles of perception. These principles served as a data base from which the items on the dimension of perception in the proposed reaction instrument were formulated.

The remainder of the items in the reaction instrument were based on the principles associated with human learning. Several of these principles, such as part versus whole learning, the motivation to learn, knowledge of results, repetition, insight, interest, the meaningfulness of material, and reward and punishment systems are enumerated in the literature and provide the information on which the rest of the reaction instrument is based. From a psychological point of view it is suggested that both the occurrence and the amount of learning in training or learning situations are dependent on the extent to which the principles of learning are applied in such situations. To phrase this somewhat differently, in the absence of such principles the probability of learning taking place, is severely reduced. It should be noted that these principles were derived, not so much from the various theories of human learning encountered in psychology, but from the repeated research results emanating from theories such as those proposed by the gestaltists, behaviourists and others.

In view of the preceding discussion the purpose of the reaction measure, stated in operational terms, can be phrased as follows: to determine the extent to which trainees perceive the principles of perception and learning to be present in a training session or programme.

The nature of the reaction measure

The terms "reaction" and "perception" as used in the previous section, need some clarification. In the context of the present study, it is asserted that they have essentially the same meaning – both point to the "experience" of trainees. The former, however, tends to be more appropriate to Kirkpatrick's model, whereas the latter has become the generic term in industrial psychology and refers to measures of "experience". Such measures, usually in questionnaire form, are designed to provide an evaluative response from individuals regarding some **external** abstract (as opposed to concrete, such as weight or sound) phenomenon.

The related term of "attitude" also needs to be distinguished. Attitude measures are designed to provide an indication of some **internal** set or readiness "to react favourably or unfavourably towards a designated class of stimuli" (Anastasi, 1976, p. 543). As such it can be viewed as a relatively stable psychological characteristic. Although the focus of attitudinal and perceptual measures may differ, it is submitted that from a psychometric point of view they are similar in terms of rationale and methodology. This suggestion was borne out by an examination of the models proposed in the literature on attitudes. Basic to many of the models (Chaiken & Stangor, 1987; Bagozzi & Burnkrant, 1979. Zajonc & Marcus, 1982 and Fishbein & Ajzen, 1975) are the "trilogy of attitudes" in one form or another, suggesting that attention be paid to the following dimensions: cognition, affect and conation. It was accepted that the reaction of trainees to the principles of learning would be a composite of cognitive, affective and conative dimensions.

The formulation of the specific items in the reaction measure required that the principles of perception and learning be translated into suitable questions so as to permit evaluation on an intensity scale. This was done so that respondents had to indicate on a six point scale, for each item in the questionnaire, the extent to which they perceived the implied principle to be present in a particular training situation. Whereas the principles of perception were merely converted into questionnaire items, the principles of learning were phrased in such a way that they had a cognitive, affective or conative appeal. In the questionnaire each of the categories (perception, cognition, affect and conation) were represented by 12 items. The questionnaire was presented to nine volunteers, whose formal education varied between 10 and 13 years. They were requested to edit the items with special emphasis on grammer, clairty and ambiguity. In the questionnaire the edited 48 items were arranged in a random order. Some examples of the questionnaire items are given in Table 1.

TABLE 1
DEFINITIONS FOR THE PROPOSED DIMENSIONS AND EXAMPLES OF QUESTIONNAIRE ITEMS

Dimension	Definition and examples of questionnaire items
Input	Definition: The ease of information transmission. Example: Given the abovementioned course to what extent did important information stand out from background information?
Emotion	Definition: Changes in awareness of internal states. Example: Given the abovementioned course to what extent did you experience positive feelings such as joy and curiosity.
Cognition	Definition: Changes in awareness of external states. Example: Given the abovementioned course to what extent were ideas and questions of your own provoked?
Conation	Definition: The urge towards purposive learning. Example: Given the abovementioned course, to what extent did you strive toward performance levels above the required minimum.

Postulates

The following postulates were formulated:

Postulate 1: In view of the data base (the principles of perception and learning) on which the questionnaire is based, it was postulated that an item analysis on each of the four categories into which the items were grouped, would yield acceptable coefficients of internal consistency.

Postulate 2: In view of the conceptual differences between the four dimensions which were used in the questionnaire, it was postulated that a factor analysis would indicate that the items in the questionnaire would indeed be measuring four different factors.

Sample

The questionnaire was administered to 362 first year students approximately four weeks prior to their examination. They were required to evaluate their training in industrial psychology. Completion of the questionnaire was voluntary and 310 participants completed the questionnaire. Of these 104 chose to remain anonymous.

RESULTS

Four separate item analyses were performed, one for each category of the questionnaire. For this purpose the responses obtained from the 310 participants were used. The results of the NP50 item analysis programme which iterated on the indices of reliability of the items, appear in Table 2. The table shows

TABLE 2
RESULTS OF ITEM ANALYSIS

Dimension	1	2	3	4
Item	$r_{gx}s_g$	$r_{gx}s_g$	$r_{gx}s_g$	$r_{gx}s_g$
1	0,60	0,57	0,94	0,75
2	0,61	0,65	0,95	0,63
3	0,81	0,62	0,92	0,66
4	0,69	0,61	0,89	0,84
5	0,87	0,73	0,85	0,76
6	0,34*	0,71	0,31*	0,84
7	0,82	0,83	0,85	0,81
8	-0,15*	0,53	0,73	0,74
9	0,66	0,64	0,81	0,58
10	0,84	0,76	0,74	0,95
11	0,71	0,72	0,89	0,67
12	0,84	0,28*	0,34*	0,82
Number of Iterations	2	2	2	1
Final criterion value	0,59	0,55	0,51	0,60

Note: * indicates items discarded during iterations process

the indices of reliability of the items, the criterion value for the final iteration, the number of iterations and the items discarded during the iteration process. The number of items (43) retained from the original 12 per category, as well as the reliability coefficients (Kuder-Richardson 20) for the categories are shown in Table 3.

TABLE 3
RELIABILITY COEFFICIENTS

Category	Number of items retained	r_{tt}
Input	10	0,81
Cognition	11	0,78
Emotion	10	0,85
Conation	12	0,84

The 43 items retained were intercorrelated, but the intercorrelation matrix is not reproduced here on account of its size. The eigenvalues of the unreduced intercorrelation matrix were computed by means of the BMDP4M programme so as to determine the number of factors present in the intercorrelation matrix. Nine factors, on the basis of eigenvalues $\geq 1,0$ (Kaiser 1961), emerged. One factor had one loading only and was discarded. The obtained factor matrix was rotated to simple structure by means of the varimax rotation. The items were accordingly categorised into eight categories in terms of their highest loadings. Eight subsets were formed by summating the items with high loadings on each factor. The eight subsets were intercorrelated (see Table 4). The eigenvalues of this matrix are given in Table 5. Only one eigenvalue was greater than unity.

TABLE 4
INTERCORRELATION MATRIX OF EIGHT SUBSETS

Factor	1	2	3	4	5	6	7	8
1	1,00							
2	0,49	1,00						
3	0,53	0,35	1,00					
4	0,54	0,42	0,44	1,00				
5	0,38	0,21	0,46	0,31	1,00			
6	0,65	0,46	0,35	0,42	0,24	1,00		
7	0,45	0,36	0,47	0,39	0,42	0,34	1,00	
8	0,41	0,29	0,40	0,32	0,20	0,23	0,19	1,00

TABLE 5
EIGENVALUES OF UNREDUCED INTERCORRELATION MATRIX IN RESPECT OF THE EIGHT SUBSETS

Factor	Eigenvalue
1	3,73
2	0,97
3	0,85
4	0,61
5	0,57
6	0,53
7	0,45
8	0,30

One factor was accordingly extracted and is given in Table 6. It thus appears that the reaction questionnaire measures only one factor.

TABLE 6
FACTOR LOADINGS IN RESPECT OF SECOND ORDER FACTOR

Factor	Loading
1	0,84
2	0,59
3	0,69
4	0,66
5	0,50
6	0,63
7	0,60
8	0,46

DISCUSSION

A number of factors, such as the homogeneity of the sample of subjects, item characteristics and test length can influence estimates of reliability. Most authors, therefore, refrain from prescribing what values are indicative of acceptable reliability coefficients. Guion (1965, p.46), however, stresses the purpose of measurement and states that: ''If the purpose is research – to test a hypothesis on a group basis – then reliability need not be high at all''.

It is envisaged that the proposed instrument would be applied initially in group settings and that the results would be used to effect changes to training situations rather than in individuals. From this perspective the size of the obtained coefficients of internal consistency can be regarded as satisfactory and supportive of Postulate 1 (as indicated in Table 3, the coefficients varied between 0,78 and 0,85 over the four dimensions). Specifically there appears to be sufficient evidence to suggest that the items within the four dimensions are homogeneous and as such would provide consistent estimates of the internal strucure of each dimension.

Contrary to the expectation stated in Postulate 2, the single factor produced by the factor analyses indicates that the items in the questionnaire are measures of the same thing (trainee reaction, for lack of a more appropriate description). The lack of support in the factor analyses for the conceptually different dimensions in the questionnaire is not novel in situations, such as attitude research, where the threefold classification of cognition, affect and conation has been investigated. As such the study thus seems to provide support for the conviction of McGuire (1969, p. 157) who remarked about the cognitive, affect and conative conceptualisation that "theorists who insist on distinguishing them should bear the burden of proving that the distinction is worthwhile".

Although the structure of the variable "the reaction of trainees" could not be illuminated in this study, both the item and factor analyses suggest that the instrument developed in this study could be used with sufficient confidence in both practical and research settings as far as the evaluation of training is concerned.

REFERENCES

Alliger, G.M., & Janak, E.A. (1989). Kirkpatrick's levels of training criteria: thirty years later. *Personnel Psychology, 42,* 331-341.

Anastasi, A. (1976). *Psychological testing (4th ed.).* London: Macmillan.

Bagozzi, R.P., Burnkrant, R.E. (1979). Attitude organization and the attitude-behaviour relationship. *Journal of Personality Social Psychology, 37,* 913-929.

Bell, J.D. & Kerr, D.L. (1987). Measuring training results: key to managerial commitment. *Training and Development Journal,* January, 70-73.

Birnbrauer, H. (1987). Evaluation techniques that work. *Training and Development Journal,* July, 53-55.

Carlisle, K.E. (1984). Why your training evaluation system doesn't work. *Training,* August, *39.*

Carnevale, A.P. & Schultz, E.R. (1990). Return on investment: accounting for training. *Training and Development Journal,* July, S1-S31.

Casio, W.F. (1987). *Applied psychology in personnel management (3rd ed.).* Englewood Cliffs, N.J.: Prentice-Hall.

Chaiken, S. & Stangor, C. (1987). Attitudes and attitude change. *Annual Review of Psychology, 38,* 575-630.

Clement, R.W. & Walker, J.W. (1979). Changing demands on the training professional. *Training & Development Journal,* March, 3-7.

Fishbein, M. & Ajzen, I. (1975) *Belief, attitude, intention and behaviour: An introduction to theory and research.* Reading Massachusetts: Addison-Wesley.

Goldstein, I.L. (1974). *Training: program development and evaluation.* California: Wadsworth.

Guion, R.M. (1965). *Personnel testing.* New York: McGraw-Hill.

Kaiser, H.F. (1961). A note on Guttman's lower bound for the number of common factors. *British Journal of Statistical Psychology, 14(1),* 1.

Kirkpatrick, D.L. (1979). Techniques for evaluating training programs. *Training and Development Journal, 33, (6),* 78-92.

McGuire, W.J. (1969). The nature of attitudes and attitude change. In G. Lindzy, & E. Aronson (Eds.), *The handbook of social psychology.* (Vol. 3). Reading: Addison-Wesley.

Morris, M. (1984). The Evaluation of training. *Industrial and Commercial Training,* March/April, 9-16.

Newstrom, J.W. (1978). Catch-22: The problems of incomplete evaluation of training. *Training and Development Journal,* November, 22-24.

Parker, B.L. (1986). Summative evaluation in training and development. *Journal of Industrial Teacher Education, 23(2),* 29-55.

Quin, S.R. & Karp, S. (1986). Developing an objective evaluation tool. *Training and Development Journal,* May, 90-92.

Smeltzer, L.R. (1979). Do you really evaluate, or just talk about it? *Training, 17(8),* 6-8.

Whitelaw, M. (1972). *The evaluation of management training- a review.* London: Institute of Personnel Management.

Wickens, C.D. (1992). *Engineering psychology and human performance.* Illinois: Harper Collins.

Zajonk, R.B. & Markus, H. (1982). Affective and cognitive factors in preference. *Journal of Consumer Research, 9,* 123-131.