35

Educational Assessment and Evaluation

BISHWA NATH MUKHERJEE

INTRODUCTION

This chapter deals with a critical review of Indian studies reported during the period 1988-1992 in the area of educational assessment and evaluation, including research in examination. The original plan was to deal mainly with studies pertaining to educational measurement, i.e., a procedure for assigning numbers (usually called scores) to a specified attribute or characteristic of persons pertaining to any teaching-learning process in such a way that the numbers describe the degree to which the persons possess the attribute. Thus, educational measurement could be broadly defined as descriptions of behaviour pertaining to any teaching-learning process that can be expressed in numbers. However, educational scores are mainly used for making decisions about students' curricula and educational policy rather than just treated as numbers. For this reason, instead of measurement, there is now a distinct preference for using the term "assessment" (Mcloughin and Lewis 1986; Jackson, D.N. and Messick, S. 1968, p. 42) which refers to a process for obtaining any kind of information (not only through mental and educational tests but also through any formal and/or informal observation, oral questioning, interview, group discussion, analysis of students' records, peer ratings, teacher evaluations, etc., or a combination of various measurements) on a person with a view to evaluation or prediction of a particular complex criterion. Educational evaluation includes value judgments concerning any educational outcomes/process in addition to a qualitative and quantitative description of this process. With the above definition as stated, academic selection, prediction of achievement, and use of qualitative methods for the study of the teaching-learning process would come under this review. Studies on educational evaluation could cover, for example, the effectiveness of instruction and the assessment of appropriateness of a curriculum or a system of examinations. The emphasis of this review is methodological rather than substantive. Generally, reports of studies attempting primarily to infer relations or test hypotheses are not cited unless they make a contribution to assessment theory or exemplify a technique of interest.

Various authors, e.g., Menzel, E.W. (1950), Harper, A.E. Jr. (1960), Mitra, S.K. (1961, 1968, 1972), Jalota S. (1965), Mitra, S.K. and Kumar, K. (1974, 1979), Mukherjee, B.N. (1979), Kulkarni S.S. and Kumar, K. (1986), Kulkarni, S.S. and Puhan, B.N. (1988), and Kumar, K. (1991) have reviewed Indian studies up to 1988 in the area of psychological and educational assessment. This report is expected to cover the major studies done in these areas till 1992. An attempt has been made here to critically review those studies and identify the various facets of the research gaps in the area of educational assessment and evaluation in India, and suggest some priority areas in the field. A few studies in the area of examination, which were reported during 1988-92, are also included here for the review.

A simple count of published studies in the area of assessment, evaluation and examinations suggests that about 500 investigations have been conducted so far in India till 1992. There are also numerous studies in these areas which are still unpublished. A few have been published in foreign journals and are not included here.

An analysis of the decade-wise distribution of publications in the above areas indicates a declining trend during the eighties and the early nineties. Only in the area of examination, there has been some increase in the number of publications. Creativity research, prediction of academic achievement, measurement of interest and values are some of the notable areas where there has been a significant downward trend so far as research productivity is concerned.

Out of the forty-nine studies, which are critically reviewed in the present report, 23 are doctoral dissertations submitted to different Indian universities, 16 are independent faculty research studies and the remaining 10 are published articles, mostly in the *Indian Educational Review* and *Experiments in Education*. The selection of these studies is to some extent biased because in the search for literature, the articles pertaining to educational assessment and evaluation which appear in other journals like *Psychology and Education*, *Psychological Studies*, Manas, *Indian Journal of Psychology*, *Indian Journal of Psychometry*, *Journal of Psychological Researches*, *Asian Journal of Psychology and Education*, and *Indian Journal of Applied Psychology* have not been included. However, references to reports published elsewhere, when encountered, were pursued if they seemed relevant.

REVIEW OF RECENT STUDIES IN ASSESSMENT

Criterion-Referenced Tests

During the period of this review (1988-92), there has been some upsurge of interest in criterion referenced tests (CRT) developed specifically to determine whether an individual has learnt specific skills or a specific knowledge or has attained a given instructional/behavioural objective. Singh, P. (1988), for example, continued his interest in CRT and developed in 1988, criterion-referenced tests in environmental studies for students in Classes III-V. Raithaththa, B.C. (1989) developed a CRT on vowel coalition *(Svarsandhi)* in Sanskrit language for Class VIII and attempted to validate it. Vaghela, V.K. (1992), similarly, developed a CRT for Class VII in social studies on the topic of "Delhi Saltnat". Although the specifications in all the above three tests describe clearly the knowledge to be measured, they do not demonstrate necessarily whether these tests are measuring the type and level of knowledge they are intended to measure. The extent of congruency of the test items with the domain definition (description validity) is, however, clear. In the case of Singh, P.'s (1988) CRT in environmental studies, there was perfect content validity as the Item Objective Congruence Index (IOCI) of each item of the remaining two tests was +1.00. However the graph-based unidimensional indices (for validity) of each behavioural domain ranged from 0.07 to 0.70 in the second case, and from 0.18 to 0.59 in the third case. The parallel form reliability in the first case was also not very high. Nevertheless, all these three CRT's are useful because simplified versions of specifications employed in these tests can be used to inform teachers and students about what will be tested and how it will be tested. The test results can also be useful for teachers who want to improve their instruction.

Ability Tests

Besides the CRT's, two ability tests have been adapted during the period under review. Panchal, D.H. (1991) and Patel, S.R. (1991), for example, adapted and standardised the first half and second half, respectively, of the British Ability Scales (BAS) for Gujarati children of urban areas. Both the investigators, working for their Ph.D. from Gujarat university, applied the Rasch (1960) model to all the adapted items of their respective scales by obtaining Item Characteristic Curves (ICC), and thereby checked the congruency of these scales with the original ones. Patel, R.S. (1989) found that out of the total 12 subscales of BAS as adapted in Gujarati, the test of verbal fluency was the only one which was not fitted to the Rasch model. The items of the remaining subscales were checked for their difficulty, ability parameters and goodness-of-fit with the model. Panchal, D.H. (1991) also found a good fit of the data of all his newly constructed items with the Rasch model. In his case, the test-retest reliability of the first half of the adapted BAS ranged from 0.63 (speed of information processing) to 0.95 (formal operational thinking). The investigations of both Panchal, D.H.. (1991) and Patel, S.R.(1991) have shown an effective use of the Rasch (1960) model in a test development programme. The model has been also effectively used in the Indian Statistical Institute, Calcutta, in a recent study on the assessment of minimum learning in primary education (Basumallik et al. 1992).

Mention should be made of another ability test which was adapted during the period under review by Kumar, A. (1990). For the try out, a total set of 76 items, 19 for Block Design, (BD), 12 for Pass Along (PA), 12 for Pattern Drawing (PD), 19 for Memory Drawing (MD) and 14 for Picture Construction (PC) was administered to 370 illiterate adults (225 males and 145 females) sampled, incidentally, from the Balasore district of Orissa. After item analysis, 48 items altogether were selected for the final form of this performance test battery out of which 12 items were for the BD, 8 for PA, 10 for PD, 12 for MD and 6 items for PC. This final form was administered to 1,020 illiterates (704 males and 316 females) in the age group 15-35 years for the purpose of norm construction. The norms were reported in percentile and T-score forms. The subjects were selected from 12 districts of Orissa on the basis of stratified random sampling.

The test-retest reliability coefficients of the different scales of the above battery were found to be 0.81 for BD, 0.83 for PA, 0.88 for PD, 0.76 for MD and 0.93 for PC. The split-half reliability coefficients were reported to be 0.93, 0.97, 0.71, 0.87 and 0.83, respectively. For the whole battery the split half reliability was 0.89. The inter subtests correlations ranged from 0.94 to 0.99 thus indicating the clear existence of a common factor which could be interpreted as "general intelligence". However, the investigator did not collect any data which could establish the construct and/or predictive validity of this performance test battery.

Reading Tests

The period under review also witnessed an upsurge of interest in the study of reading. This feature is quite encouraging in view of the fact that there are so few tests in India in the area of reading comprehension and writing ability. Chawla, S. (1988) suggested a general procedure of constructing a test of reading comprehension using multiple-choice items. She, however, failed to give any operational definition of reading comprehension and to distinguish between a survey of reading skills and a diagnostic test of reading skills. It appears that Chawla, S.'s (1988) paper deals with the steps involved in designing a multiple choice test which can estimate the status (survey) of one's reading comprehension rather than a diagnostic tool which can help in identifying the errors involved in: (a) word recognition, (b) syllabication, (c) sound discrimination, and (d) oral reading errors such as omissions, reversals, mispronunciation, etc. The paper illustrates the conceptual defects arising out of the author's neglect to operationally define the construct of reading comprehension and to use any kind of theoretical perspective.

Raviya, D.L. (1990) constructed a reading comprehension test in the subject of Sanskrit for the eighth class students.of Saurashtra region. For the purpose of standardisation and norm construction, 3,725 sample students (1,441 boys and 2,284 girls) were chosen by stratified random sampling. Norms were prepared separately for boys and girls in terms of percentile rank, Z-scores and T-scores. The test may be used for survey of reading comprehension in Sanskrit for Class VIII students, although its validity is not clearly established.

Neither Chawla, S. (1988) nor Raviya, D.L., (1990) made use of Royer's Sentence Verification Technique (SVT) for measuring comprehension. A considerable amount of research has been done on the SVT during the last fifteen years and it has been shown that the technique is especially useful in measuring passage comprehension. Even Mukherjee D.P.'s (1991) comparative study, cited later in his review, also failed to include the SVT.

Patel B.V. (1988) of Sardar Patel University attempted to construct and standardise a tool for measuring reading comprehension and speed in Gujarati for students of Classes V to VII, which could be used for evaluating the effect of a reading improvement programme. These authors also failed to use the SVT. The reliability and validity of the test were not assessed by the author. However, by using two equated groups —experimental and control—they found the test useful in demonstrating the positive effect of a reading improvement programme. Students with a high SES level who took the experimental treatment made greater improvement than students of the same experimental group with a low SES level. This finding can be construed as a construct validity of the test.

The concurrent validity of the Patel-Vora (1988) Reading Comprehension Test is also borne out from the doctoral dissertation project of Dave, M. (1992) carried out in Gujarat University, Ahmedabad. The correlation between the scores on the reading speed showed significant relationship with intelligence, vocabulary, achievement motivation and study habits.

A test of reading readiness in English for pre-school children has also been constructed during the review period by Manjula, R. (1991). Reading readiness tests are devised mainly to identify children who are not psychologically prepared at the prescribed time to benefit from instruction, although not necessarily retarded in general mental development. Manjula R.'s (1991) battery consists of seven subtests, namely, auditory memory, letter recognition, sound-letter correspondence, visual matching, letter pattern, listening and quantitative concepts. Each subtest included only 20 items. No attempt has been made to report the results concerning the reliability and validity of the proposed test which the author plans to standardise for the first graders who are to be admitted in the schools of Bangalore city. Although it is generally difficult to determine the validity of a reading readiness test for pre-school children because its effectiveness as a predictor has to be estimated in terms of progress at the earlier stage, certain predictive studies could have been undertaken to forecast achievement in reading in Classes I and II. It is expected that follow-up studies would be undertaken in the future with Manjula's (1991) reading readiness test.

A methodological study was reported by Mukherjee, D.P. (1991) which dealt with the comparative effectiveness of the free-response type of items (cloze type) and multiple-choice items for measuring reading comprehension at the eighth class level in West Bengal. Mukherjee, D.P. (1991) came to the conclusion that the cloze-type items facilitate diagnosis of different types of errors involved in comprehension of printed passages which can be useful for language teachers and textbook writers. The author, however, did not extend his study to demonstrate the relative superiority of the two techniques of item writing in terms of different psychometric properties and prediction of scholastic achievement.

Ashai, Y. and Mohite, P. (1989) made an attempt to standardise a teacher's rating scale for identifying children with learning difficulties in reading and writing. For this purpose, a sample of 720 children was selected through stratified random sampling from Gujarati-medium schools in Baroda. Norms were prepared using percentile ranks. The sample size, however, appears to be too small for proper comparisons. Methodologically, the study seems to be quite weak.

A test of reading and writing skills for pre-primary school children has been constructed by Singh, N. (1988). Since the test is based only on 200 students from Delhi schools, and no data have been presented with respect to the reliability and validity of the test, the effectiveness of the test cannot be assessed. The study also suffers from other methodological and conceptual defects.

Achievement Tests

A few achievement tests, mainly in science, English and Marathi, have been constructed and standardised during the period under review. Rozario, L. (1989) constructed and standardised achievement tests in physics, chemistry and biology for students of Classes VIII and IX studying through the English medium in Bombay suburban schools. These tests were found useful in identifying low achievers in each of the above three branches of science and in analysing their specific learning difficulties.

Kulkar, K.R. (1989) used a random sample of 651 students of Class VIII drawn from 16 Marathi-medium secondary schools of Kolhapur, Maharashtra, to develop five unit tests in Marathi. The standardisation of the tests was done on a sample of 5,355 pupils from 23 secondary schools in Kolhapur. The Kuder-Richardson reliability of the five tests ranged from 0.775 to 0.964. These five tests cover only five units from the Class VII Marathi textbook. Therefore, their practical utility is very limited.

Mention has already been made of the achievement tests in Sanskrit and social studies constructed, respectively, by Raithaththa, B.C. (1989) and Vaghela, V.K. (1992) using the criterion-reference approach. The reliability of the mastery-non-mastery classification decision ranged from 0.79 to 0.90 in the case of Sanskrit vowel-coalition test, and from 0.91 to 0.96 in the case of social studies. Therefore, both the tests appear to be reliable for mastery-non-mastery classification. Although useful for the purpose of diagnosis of low achievers and their difficulties, these tests cannot be used at present in guidance and placement unless their "predictive" validity is demonstrated.

A language ability test in Hindi as a subject was constructed and standardised by Vyas, S.G. (1988) for pupils of Class XI of the Saurashtra region. A total of 1,677 students (993 boys and 684 girls) belonging to both urban and rural areas were selected for this purpose and both T-scores and percentile ranks were obtained as norms. The test is norm-referenced rather than criterion-referenced and as with any other norm-referenced achievement test, it does not indicate what the students know or do not know. The results obtained with the Vyas Hindi ability test give teachers and policy-makers very little information for improving schools or for undertaking a remedial teaching programme.

It appears from the above citations that during the period under review, there has not been any work on the construction and standardisation of tests in different regional languages, other than Marathi and Hindi. The previous work on the construction of achievement tests in most regional languages, except possibly in Gujarati, has been quite sketchy and fragmentary. For this reason, this reviewer totally agrees with Singh, P. and Prakash, V. (1991) in emphasising the need for undertaking systematically the construction and standardisation of achievement tests at the national as well as regional levels. However, the new tests must be technically well planned and meticulously constructed. The normative data should be based upon the scores of many thousands of students and not just a few hundred, as in the case of most of the available Indian tests of achievement. In addition to content validity, their construct and predictive validities must be demonstrated after a rigorous item analysis.

Miscellaneous Tests

In addition to the tests discussed above, ten more tests and scales have been constructed during the period under review. Mention may be made of Patnaik, S.P.'s (1990) study which addressed the problem of developing and standardising a situational test for measuring teaching aptitude appropriate for elementary level. The 54-item situational test developed by Patnaik, S.P. (1990) was intended to measure: (a) teacher competence, (b) general teaching method, (c) teaching skills, (d) teacher behaviour and motivation, (e) remedial teaching and evaluation, (f) relation with home, parents and community, (g) attitudinal characteristics and personality, and (h) physical and organisational ability. The sample used for the purpose of construction and standardisation consisted of 915 teachers working in elementary schools in Orissa. The split-half reliability was found to be 0.47, which is considered to be low. The teachers with 5-10 years of teaching experience and with postgraduate qualifications generally scored higher on this situational test than other groups of teachers. This result only shows that the test may be useful for evaluating professional knowledge in teaching rather than aptitude for teaching. Other types of validities, particularly predictive validity, must be established before one asserts the usefulness of Patnaik, S.P.'s (1990) test for selection of teachers. From the low split-half reliability of the test, it seems doubtful, however, whether rigorous item-analysis was done before item selection. It is also necessary for the test constructor of any situational test to show that the selected test items really demand activity commensurate with the situation that simulates the task to be performed (here teaching). The face validity of each item of Patnaik, S.P.'s (1990) situational test of teaching ability should have been demonstrated. However, the effort on the part of the test constructor to develop a situational test is commendable. In the area of Indian psycho-educational assessment, there is a dearth of situational tests which can approximate work samples.

Another aptitude test which needs some mention is that of Sharma, S.K. (1991) who made an attempt to develop a battery of tests of scientific aptitude for the students of Class XI. The items for the different tests, namely, scientific awareness, numerical ability, perceptual ability, mechanical comprehension, reasoning ability, spatial ability and figure dexterity, were tried out on a sample of 400 students of Jammu City after a comprehensive "content analysis'. Unfortunately, the writer has not reported any of the studies pertaining to reliability and validity. A good test of scientific aptitude is expected to estimate the probability of success in scientific and engineering occupations. Sharma, S.K. (1991) has not been able to show this good feature with respect to the test he has devised. He also failed to establish the desirable psychometric properties of the test. Therefore, the study is quite weak in terms of its technical soundness.

Creative thinking is directed toward the production of a new form—new in the sense that the thinker is generally not aware of the form before he/she begins the particular line of thought. Creative thinking can be used to encourage conceptual development as in the discovery method of teaching. For this reason, measurement of creative thinking is regarded as an important area of research. In view of this, Siddiqui, S. (1988) undertook the task of constructing and standardising a test of creative thinking for Urdu-speaking students in the Telangana area of Andhra Pradesh. A battery of non-verbal tests, namely, Picture Construction, Incomplete Figure, Repeated Figure activity, Consequences, Novel Uses, Similarities, Product Improvement and Story Title was administered to 1,000 boys and 1,000 girls who were studying in Classes VIII, IX and X of secondary schools in urban and semi-urban areas. Although the author claims to have selected the items on the basis of item-validity, no external criterion was employed as an indicator of creative thinking. It is doubtful if the proposed test of Siddiqui, S. (1988) really measures one's ability to define, think or perceive in a way different from the established or usual way.

Chauhan, R.'s (1988) academic alienation scale also suffers from methodological defects. As in the case of Siddiqui, S. (1988), no attempt has been made to define operationally the construct under study. Not a single coefficient of reliability and validity is reported. The group differences, as reported by Chauhan, R. (1988), do not reflect the validity of the subtests measuring academic cynicism and academic inefficacy. No attempt has also been made by Chauhan, R. (1988) to asses the role of social desirability on the total alienation scores. As is true with most personality inventories developed in India, Chauhan, R's (1988) scale is an example of a high-sounding but loosely defined psychological construct in total disregard of the response bias. Such attempts are considered to be a conspicuous waste of research.

The projective inventory approach to personality assessment makes use of structured projective situations and is expected to minimise, if not completely eliminate, the social desirability (SD) contribution to personality-scale scores (Puhan, B.N. 1982). It seems, therefore, that Behara, N. (1990) who attempted to standardise a projective inventory for measuring adjustments in the areas of health, home, social and emotional well-being, should have studied the role of the SD factor. This precisely was not done by Behara,N. (1990) although he demonstrated satisfactory reliability of the test scores.

The role of social desirability was also not assessed in the case of Shah, J.H.'s (1989) Self-Concept Inventory (SCI) for Gujarati students of Classes XI and X. The test is made up of 80 adjectives which could be rated on a 5 point scale for appropriateness of self-description and was standardised on a sample of 718 students. In the case of the adjective check-list, the respondent marks all the adjectives he/she considers to be descriptive of himself/herself. In the case of the SCI, all the adjectives are to be rated in terms of their appropriateness of description. Thus, the SCI adjectives are more like unrestricted Q-sort items. In all these measures of self-concept, the role of SD must be assessed independently. Otherwise, the scores obtained on the self-concept measures can be interpreted as the respondent's predisposition to make a favourable social impression about his/her personality. The three correlational studies reported by Shah, J.H. (1989) in this report neither rule out the above interpretation of her scale nor establish any concurrent or construct validity of the SCI.

The author also did not address some of the difficult conceptual and methodological problems involved in the measurement of self-concept. No operational definition of self-concept was given. It is not necessary, however, to define something in order to study it empirically, so long as its assumption provides a useful conceptual link in the establishment of lawful relationships with other things. Milgram N.A. and Helper M.M. (1961) demonstrated long ago one of the great difficulties of doing this in the study of self-concept by an experiment which shows that social desirability, in particular, and response sets in general, are powerful effects on self-concept inventory scores or on any kind of self-ratings. Thus, unless more studies are undertaken, it is difficult to say if Shah, J.H.'s (1989) SCI scores are reflections of positive self-esteem or of the ability to fake weil or the operation of both faking and self-esteem.

The question of social desirability response set does not arise in the case of the Art Judgment Test (AJT) as devised by Ambasana, A.D. (1989) but other methodological problems, such as whether the test is in large part a measure of learning rather than aptitude, are not tackled. It is also not clear whether the test taps perceptual facility, creative-imagination, aesthetic intelligence and aesthetic judgment, or a combination of these. The author has not taken the trouble of undertaking any concurrent validity study for correlating the art judgment scores on such measures as the Mier Art Judgment Test, the Knauber Art Ability Test, the Horn Art Aptitude Inventory and the Graves Design Judgment Test. Neither was any attempt made to correlate the obtained scores on the AJT with art grades and/or ratings of artistic ability of the respondents.

The last scale to be included in this section is the one developed by Rao, R.R.S.P. (1991) for measuring people's attitude toward the new educational pattern. The scale was developed by employing the Thurstone five-point equal appearing interval scaling method. The attitude scale was administered randomly and individually to 1,000 individuals broadly grouped as students, teachers, lecturers, lawyers, businessmen and administrators to assess their attitude towards the nature of curriculum, methods of teaching, teaching aids, the requirement of doing socially useful productive work, evaluation and examinations, vocationalisation, and science and mathematics education. No clear description is obtained about the people who were asked to judge to make objective evaluations of the positions of the items on the attitude continuum. A frequent criticism of the Thurstonian scaling method is that the characteristics and attitudes of the people who judge the items may be very different from those of the respondents whose attitudes are to be scaled. The population of judges also may not be homogeneous. In the absence of any description about these and also about the goodness of fit of the item data to the non-monotone probability model connected with the Thurstone scaling procedure, it is difficult to assess the appropriateness of the method for measuring people's attitude toward the new educational system. However, on the basis of his empirical study, Rao, R.R.S.P. (1991) concluded that the Likert type of scale was more sensitive as compared to the Thurstone scale for measuring the intensity of different attitudes. This conclusion is obvious in view of the fact that it is generally very difficult to construct items using the Thurstonian method that represent higher levels of both positive and negative attitudes.

Trends, Gaps and Priority Areas in Assessment

From the earlier review of studies in the area of educational assessment in India as well as from the detailed commentaries given above regarding the tests developed during 1988-92, certain disturbing trends are visible. As in other areas of educational research in India, there has hardly been any attempt at systematisation, integration and follow-up of the studies conducted so far. Barring a few stray cases, most of the educational tests have been virtually developed on the basis of a 'one-shot' study. The use of the one-shot study completely ignores the temporal variables and leads to serious internal validity problems, especially for the hypothesis testing research involved in construct-validity study. Research in the area of psycho-educational assessment in India has not been so far cumulative in nature. One gets a dismal picture of the state-of-the-art of the psycho-educational assessment procedures currently in vogue in India. Ironically, there are quite a few tests and tools to screen, identify and measure a broad range of child, adolescent, student and teacher behaviour. However, because of a faulty system of conceptualisation, indiscriminate borrowing from the West and scant regard for estimating different types of imprecision in the tool, much of what has been done so far in the field is almost useless. Mainly because of defective operationalisation and measurement, the trend of which is still continuing as is evident from the discussion appearing in the above section, studies conducted in India in the area of test construction have been trivial and many of them are methodologically deficient. This is so because most studies do not show much rigour in terms of precision of measurement, in drawing appropriate inferences regarding the constructs which the researcher intended to measure and also in terms of their accuracy of generalisation, besides the negligence of a proper conceptualisation (Mukherjee, B.N. 1979, p.4). Largely because of defective conceptualisations, operationalisation and measurement, as is also evident from the recent studies reported during 1988-92, the use of such defective tests and tools has made a large portion of educational research in India quite redundant. Studies emerging from the use of such defective tools call for immediate quality control in educational research (Mukherjee, B.N. 1992).

It is not very difficult to see in most Indian psycho-educational assessment literature a total neglect of proper conceptualisation and theoretical underpinning as well as absence of rigour in methodology. For this reason, the whole field of psycho-educational assessment in India is almost virgin and there is a need for starting from scratch if we are at all serious about indigenisation.

Previous studies in the area as well as the ones reviewed here have generally failed to address specific issues which have relevance or practical implications connected with the teaching-learning process. This unfortunate trend of generally not pursuing any issue-based studies is still continuing. Seldom do we find any serious effort in India to make educational testing a policy tool to improve teaching and learning.

As an example, mention may be made of the recent studies on reading comprehension reviewed above or the ones conducted earlier. Not a single one of these deal with theoretical issues involved in beginning reading instruction such as the relative effectiveness of 'meaning first versus discrimination first' approaches (Harris, T.L. 1962) or a combination of the two which emphasises learning of differentiated wholes in contrast to chiefly wholes or chiefly parts.

Educators and test developers in developed countries have started redesigning tests to influence the teaching-learning process in ways that are consistent not only with good teaching practice but also with new viewpoints about how students make use of their time, and learn, and solve problems. Unfortunately, this trend has still to appear in India. Few educational researchers in India have so far taken a bold initiative to design as well as redesign and use a technically accurate, educationally sound and intellectually thoughtful assessment and evaluation programme.

Another noticeable feature in Indian educational assessment literature is the neglect to use a relevant theoretical framework in a test development programme. Only in recent years, the Rasch model and Lord's model have been used by a few Indian researchers, such as Natarajan, V. (1984), Panchal, D.H. (1991), Patel, S.R. (1991) and Basumallik, T. et al. (1992). Very recently, the two parameter logistic model has been used to obtain minimum expected score as an approximate measure of minimum learning in Bengali and mathematics for primary schoolchildren at the end of class IV in 17 districts of West Bengal in a study jointly carried out by ISI and SCERT (Roy, Guha, Mitra and Roy, 1995).

The Guilford SI model as well as Piagetion theory have also served only in few cases [Khire, U. (1989); Rani, R. (1993); Shashilatha (1977); Sandhu, (1981); Jain, S.C. (1982); Padmini, T. (1980); Dash, U.N. and Das, J.P. (1984).] Other than these, the majority of educational researchers have seldom devoted thought to supporting their assessment procedures with theoretical and empirical justifications. This unfortunate trend of conducting research without using any relevant theoretical framework or conceptual model is still continuing in India in spite of a few critics' repeated notes of caution in this regard (Nandy, A. 1976; Mukherjee, B.N. 1979; 1993a; Kumar, K. 1991; Sinha, D. 1986).

Mukherjee, B.N. (1993a) has suggested the possibility of using a number of theoretical frameworks, including the behaviour-ecological model and the information processing approach, which can help not only to guide test development research in general but also help in determining the construct validity of the tests or instruments. A framework for achievement testing originally proposed by Shoemaker, D.M. (1975) which points to the need for item-banking and matrix sampling, for example, might be quite fruitfully used for construction of proficiency tests in India.

The trend of constructing educational tests and tools which have a very high urban bias is also continuing as will be evident from the studies reported by Manjula R. (1991); Ashai, Y. and Mohite, P. (1989); Singh, N. (1988), Sharma, S.K. (1991); Vaghela, V.K (1992); Panchal, D.H. (1991) and Patel, S.R. (1991), all of which are reviewed in the previous sections. Although a large number of available educational tests have grade norm, age norm and percentile ranks for interpretative purposes, they are mostly usable with local school-children studying in urban areas. The test items are generally so constructed that they do not have much relevance for rural children. The so-called standardisation data (often based on a small-sized sample) are also obtained from children already attending schools in urban areas because it is convenient to reach them.

By and large, researchers in the field of educational assessment are still not showing any painstaking effort in studying the quality of their data and identifying the various factors influencing the imprecision of their instruments. The quality of data is governed by its reliability, accuracy, completeness, comprehensiveness and timeliness. If the tool or instrument used for data gathering is technically not sound, then it would be difficult to achieve these desirable attributes. As a matter of fact, very few educational tests have been developed so far in India with a good statistical design ensuring proper sampling of both items and individuals. Very few educational researchers in India have constructed their instruments with a given logical design so that using the analysis of variance, they can triangulate their data and identify the various factors underlying test unreliability. Guttman, L. (1954); Foa, U.G. (1965); Guilford, J.P. (1965) as well as Bock, R.D. et al. (1969) have fully made use of this rational strategy in test construction. When such rational strategies are used for developing different types of educational assessment procedures, the analysis of covariance structures (ACOVS) of the test data serves as an important statistical device. The procedure not only enables us to estimate the role of different sources of variations in the data but also the extent to which the results concerning the structure of interrelationship among the components of the tool as expected from the theoretical design correspond with the observed (sample) results. Bock, Dickens and Van Pelt (1969) for example, employed the ACOVS in estimating the different components of variance in assessing the content-acquiescence correlation in the MMPI. Bramble W.J. and Wiley, D.E. (1974) used the ACOVS for investigating content variance, variance due to non-content characteristics of items, and the covariance, of content and different item characteristics.

The above applications show promise of ACOVS as a powerful technique for investigating the structure of psychological and educational tests. "If the test items are constructed systematically to represent the construct(s) being measured and various item characteristics of interest, the variance and covariances due to these sources may be estimated and assessed. For example, a test may be reformulated or constructed by each source, as well as the covariation among sources could be assessed using the covariance structure procedure" (Bramble, W.J. and Wiley, D.E. 1974, p.189). In addition to estimating various types of reliabilities, the procedure is very much suitable for obtaining an overall index of the convergent-discriminant validities (Campbell and Fiske 1959) or the validities emerging from multi-trait multi-method matrices.

Mukherjee, B.N. (1973, 1976) has shown the usefulness of ACOVS as: (a) a measurement tool for developing reliable scales, indicators and test batteries, (b) a procedure for examining many forms of validity of the tools, especially the ones devised from the point of view of multi-trait multi-method framework, (c) a methodology for testing structural hypotheses and evaluation of different types of covariance structure models, (d) a technique for theory testing after integrating the assumption of measurement errors associated with tools used for measuring the central concepts under study, (e) a statistical technique for making causal analysis with both recursive and nonrecursive equations. Thus, the data of all tests developed employing the facet design (Guttman, L. 1954 Foa, U.G. 1965; Canter, D. 1985) or the Guilford Structure of Intellect model can be subjected to ACOVS, and if the expected pattern of relationships among a set of items (constructed from a logical system of classification) is found tenable, this finding itself can be interpreted as a strong evidence of the construct validity of the test. Boruch R.F. and Wolin L.'s (1970) paper on the multi-trait-multi-method approach demonstrates the usefulness of ACOVS for estimating trait, method and error variances attributable to a measure designed from the facet point of view and the generalisability theory of Cronbach L.J. et al. (1972).

Specific structural models have been proposed for intellect (Guttman, L. 1965; Guilford, J.P. 1965), achievement and analytical ability test (Guttman, R. and Schlesinger, I.M. 1967), academic and non-academic self-concepts (Marsh, H.W. 1987), and many other educational and psychological concepts like achievement value (Mukherjee, B.N. 1974 b), persistence (Mukherjee, B.N. 1974 a), etc. Such investigation of patterns and structures, according to Murphy, G. (1969), is one of the emerging trends of psychology in the year 2000. Guttman, L. (1971) conceived educational and psychological measurement as structural theory. The whole field of mathematics has been treated as the study of patterns and structures by Rucker, G. (1987) in his famous book *Mind Tools*. In this book, Rucker discusses the 'new archetype' of patterns of mathemtics-information. Foa, U.G. and Turner, J.L.. (1970) also believed that "substantive and methodological progress will be toward an increasing interest in the study of structures" (p. 246). According to them, "experimental work on structural changes and longitudinal studies of structural growth will provide the knowledge required for devising special remedial and preventive methods" (p. 245). This trend in the study of structure and pattern is almost completely lacking on the Indian educational measurement scene even though Mukherjee, B.N. (1966) emphasised its usefulness in the field of psycho-educational assessment almost thirty years back.

There are quite a few priority areas of research in the field of educational assessment. Mitra, S.K. (1972) more than twenty years ago listed some priority areas of research in the *First Survey of Psychology in India* in the context of testing and methodology. Not much systematic work has been done in any of these areas. For example, a few test constructors report the convergent discriminant validity indices, the result concerning distractors in multiple-choice items, rigorous use of sampling techniques, and experimental studies of tests, testee and situation variables in influencing the scores on any educational tests. Educational researches in India still continue to evaluate predictors simply by testing the statistical significance of correlation coefficients. They frequently forget that statistical significance does not necessarily imply practical significance. A validity coefficient of 0.05, for example, based on as many as 5,000 respondents, may be statistically significant at the 0.01 level but it may have very little practical utility. Moreover, the style of reporting a single validity coefficient (or a single reliability coefficient) which is still continuing in India is not very meaningful even for evaluating the psychometric property of the tool or instrument.

Reliability is situation-specific. Therefore, different types of reliability coefficients (e.g. test-retest, Cronbach Alpha, odd-even, parallel form, homogeneity coefficients, stability coefficients, etc.) should be provided. In addition, the results of a component-of-variance analysis should be reported for estimating the various sources of errors or imprecision in the instrument. Similarly, in addition to discriminant and convergent validities, education researchers should report the results of construct validation, predictive and/or forecasting efficiency, synthetic validity, factorial validity, cross-validations and the utility obtained through using the particular instrument *vis-a-vis* other readily available tools and random decision-making. For example, the efficacy of different college admission tests over and above the teachers' ratings and percentage of marks obtained in the high school examination should be reported to assess the worth of the admission tests. Similarly, "the extent to which the administration of the test or questionnaire and the measurement process affecting the person being measured must also be assessed in addition to estimating the magnitude of response bias, faking, etc. Then the data could be properly assessed in terms of its potentiality for generalisability" (Mukherjee, B.N. 1985, p.181).

Although a few educational researchers in India have shown their meticulousness in establishing the reliability and validity of the instruments they have developed, almost all of them regard the reliability and validity as the *sine qua non* of any educational and psychological measurement. It is true that without attaining satisfactory consistence and accuracy, data gathering would be analogous to measuring with a rubber stick. Still, reliability and validity are not the only desirable properties of any instrument. In addition to reliability and validity of the data, we need to pay attention to the necessity of adequate sampling (both testee and test items) in collecting the data and to the objectivity and general utility of the data gathered through the instrument in making correct decisions. If the data gathered with the help of the instrument or test are too costly in terms of cash outlay in their collection, or time in scoring or editing, the utility of such data and the worth of the tools used for the data gathering are curtailed markedly. Thus, a test of low reliability and validity would be more likely to be retained according to modern decision theory (Cronbach L.J. and Gleser, G. 1966), if it were short, inexpensive, adapted for group administration and easy to score. An individual test calling for the service of a trained tester or clinician and expensive equipment would need a higher validity to justify its retention as a tool for decision-making.

Contrary to the classical psychometric theory emphasising the need for very high reliability and validity of the test, there is an increasing recognition now in all developed countries of the need for greater coverage or bandwidth at the cost of lowered dependability. But this trend of adopting the decision-theoretic approach in test construction has still to emerge in our country. Very few test constructors in India would agree with Ebel, R.L. (1961) that it is not absolutely necessary that all tests and instruments must have high validity. But this is certainly true of tests if the data obtained from their use are otherwise meaningful and helpful in making correct decisions in the area of admission, selection, classification and promotion. Cronbach, L.J. and Gleser, G. (1966) observed in this connection that "among the important wide-band procedures are the interview, the projective technique, the essay examinations and analysis of patterns of success and failures on ability tests. Each of these wide-band devices is unsatisfactory, by the usual standards of predictive efficiency and reliability. Our work suggests, however, that the negative evidence may not bear on the usefulness of the procedures for the functions they best fulfill" (p.144).

Very few test constructors in India have been inspired by the decision-theoretic approach. Most of them still follow the conventional approach and have not kept themselves abreast with Cronbach, L.J. et al.'s (1972) work on generalisability theory. According to Kulkarni, S.S. and Puhan, B.N. (1988), test developers in India are still obsessed with the classical notions of consistency as well as accuracy and "struggle for reporting high reliabilities or validities without understanding their nature and conditions under which they are calculated" (p. 67). This unfortunate trend must change if test developers in India are really interested in making significant contributions to the field.

Instead of the conventional psychometric interest of reliability and validity, we now need to be more concerned with the value or the utility of the test in arriving at different decisions regarding the examinee or the school, including a statement of the gain or loss in making such decisions. Cooil, B. and Rust, R.T. (1994) have proposed a decision-theoretic approach to evaluate "reliability as the 'proportional reduction in loss' (PPL) that is attained in a sample by an optimal estimator". Educational assessment should be able to provide the teacher and the counsellor with such information and other data that must be of value in bringing about substantial improvement in our human development programme. We must now start to judge the worth of any test more in terms of the value which the users find in it in the diagnosis and/or planning of curriculum/ instructional interventions and/or remedial programme rather than the one or two statistically significant validity coefficients. This conclusion is all the more pertinent in India when judged in the context of the emerging trend in the developed countries toward ecologically based assessment, measurement using an interactive and adaptive approach, a system-based approach or a combination of these. Unfortunately, not a single study has been undertaken so far in India to evaluate the effectiveness of psycho-educational assessment procedures in terms of their value for diagnosis, prognosis, screening, placement, intervention and usefulness with regard to decisions concerning learning outcomes about the testee and/or the group taking the test. From a decision-theoretic point of view, this is an important research gap.

In addition to the gaps identified by Mohan, S.; Pant, D. et al. (1992), it may be observed that we do not have so far any school-based satisfactory measurement procedures of those factors that interfere with, disrupt or are incompatible with the educational achievements and social development of the student, in general, and rural as well as emotionally disturbed children in particular. We need also satisfactory educational tests for out-of school children and neo-literates. We still have not developed proper tools for the evaluation and monitoring of our non-formal education programme. In the backdrop of national priorities, we need to develop technically sound tests and tools for assessing different levels of knowledge of the three R's, tests for selection of students in different vocational trades, scales for primary mental abilities that have a minimum urban bias, tests for mentally and physically handicapped children for their training and placement and tests for the assessment of behaviour disorders in the school setting. We also need to develop and standardise tests for measuring critical thinking which in recent years has become the key for educational reform because of the needed shift from didactic learning to thinking skill.

The principles of educational measurement are seldom enumerated in India before constructing any educational assessment procedure and since the effectiveness of any educational test depends to a large extent on the soundness of the principle of measurement underlying its construction, it is doubtful if most of the tests constructed are effective. There has been also no attempt in laying the mathematical, statistical and conceptual foundations of educational and psychological assessment procedures. During the period under review, only one statistically oriented paper addressed to the problem of setting effect size was published by Cruise, R.J. (1988). Although the author made it clear that the effect-size of any research is related to the meaningfulness of the true difference whereas the alpha level of significance is related only to the status of the observed difference, the idea is not new (Bakan, D. 1966; Glass, G.V. and Hakstian, A.R. 1969; Dodd, D.H. and Schultz, R.F.1973; Dawyer, J.H. 1974); K. Veren, G. and Lewis, C. (1979), for example, has recommended the use of omega square as a measure of size effect for fixed effects factorial analysis of variance (ANOVA) experimental design.

Educational researchers with a good training in mathematics have also not shown their concern in the development of sound procedures in areas involving inexact measurement. The less tangible areas of measurement are hardly receiving any attention.

Mukherjee, B.N. (1993 a) has listed the following seven broad areas of research gap in the area of educational assessment. These are:

1. Research is urgently needed for reconceptualisation of concepts like achievement, intelligence, behaviour disorder, learning disability, value, mental retardation, adjustment, creativity, teaching effectiveness, critical thinking, and many such terms. Most of these constructs are inherently multidimensional in character. Yet, the definitions available mostly from Western scholars are unidimensional in nature and may not be very relevant and valid in our culture. Certain concepts like learning disability, emotional disturbance, mental health, adjustment, intrinsic motivation, teaching competence, etc., cannot be uniformly defined even within a single culture because of different expectations of behaviour established by different groups. Behaviour that is adaptive and functional in one culture may be considered disturbed in another. The term "culture-free" test is a misnomer even within a single culture. Therefore, any test adapted from foreign sources must be first analysed in terms of the concept or concepts used in it and how relevant those concepts are for the Indian conditions, in general, and different sub-groups in particular. According to the need of the situation, proper steps should be then taken to reconceptualise the construct and make necessary adaptation for developing new tests. It will also be necessary to formulate a multifaceted conceptualisation of the construct involved in the testing and use multivariate procedures for the test development programme. When a multifaceted approach (e.g., multi-trait, multi-method) to assessment is followed, each measure provides some validity check on the others. When two or more measures of the same characteristics or attributes are in disagreement, the researcher has an indication that at least one of the measures should be questioned, modified or excluded.
2. Selection of relevant theoretical frameworks and their modification in the Indian context.
3. Research for building up a large pool of items based on a depth interview of knowledgeable persons, case studies and document

analysis.

4. Item analysis research for item calibration and selection.
5. Multivariate techniques for establishing psychometric properties of the tests.
6. Standardisation research.
7. Cost-effectiveness studies of various assessment procedures.

In addition to these research gaps, each of which is discussed at length in Mukherjee, B.N. (1993a, pp. 86-96), there is also an urgent need for using both the qualitative and quantitative approaches to educational assessment. "For enriching educational research, we need both analytical and synthetic orientations, micro and macro data, statistical and clinical predictions, deductive and inductive inferences as well theoretic objectivity and practical valuations. The unification is expected to pave the way for more precision, objectivity as also synthesis in understanding different educational phenomena (Mukherjee, B.N. 1993b, p.383). The proposed unification will also make it possible for studying both the process and the product, structure and function and, most importantly, the otherwise neglected "interactions between organismic and treatment variables" (Cronbach, L.J. 1957 p.682). Although the proposed list of priority areas is not exhaustive, mention must be made of the needed studies in the development of adequate criteria discussed separately in the following section in view of its importance.

Criterion Research

One of the main reasons why technique research connected with the development of various tests and tools useful for educational decisions has not been satisfactory so far in India is the failure of researchers to deal adequately with the problems of criteria and validity. In the area of creativity, for instance, the most difficult and important problem has been to isolate and quantify meaningful criteria of creativity. However, almost no systematic attempt has been made to study the intuitive aspects of judgment involved in rating the quality of creative products. Similarly, in the field of teacher effectiveness, not much effort has been made to investigate the factor influencing raters' judgment of teaching quality. In our search for the criteria in evaluation of instruction, we have seldom gone beyond the intended outcomes. We need to look for all possible consequences of the teaching-learning process and not just the intended outcomes (Messick, S. 1970). Very little attention has also been paid to problems of assigning values to different types of learning outcomes. Even attempts to check the goodness of a proposed criterion in terms of its reliability, validity, acceptability, predictability and other attributes as listed by Bellows, R.M. (1941) have not been systematically done. The sources of value judgment such as those made by teachers, peers, parents, supervisors, academicians and the nature as well as the extent of variations in their value judgment have also not been studied properly. The competence of these sources to give candid and honest judgment, their qualifications, experience and opportunities for observations, skill and articulation in making judgment have also not been subjected to empirical enquiry.

In addition to the above mentioned aspects, immediate attention should be given to the methodological aspects of criterion research such as sampling, statistical analysis, control of variables, assessment of psychometric properties, the time dimension of criteria, the multiple nature of the criterion, adequacy of record keeping and utilisation, minimisation of halo and other non-sampling errors and subjective bias in evaluation. When assessing the possibility of obtaining accurate self-reports from teachers, for instance, due care should be to taken to study the role of social desirability and other response sets. Similarly, when obtaining students' rating of their teachers' quality of teaching, the researcher must be on his/her guard to rule out the "score satisfaction" hypothesis. This pertains to the tendency on the part of the students to give more favourable teacher evaluation not because of a high quality of instruction but because they have done well in the relevant examination. Different types of reliabilities and validities must also be established for the criterion which in most cases will be multidimensional in nature. In creativity research, for example, we are required to study the psychometric properties of the quality, quantity and breadth of applicability of the product. A thorough study of the product variable is necessary because it represents an objective, tangible event upon which one can anchor inferred constructs such as creativity, teaching style, teacher effectiveness, critical thinking, mastery-non-mastery, leadership, etc., of direct psychological and educational interests. In measuring the quality of the product, the researcher must keep in mind that the methods used to assess quality (i.e. subjective checklist, rating scales, ranking methods) are often unreliable and subject to changing fads and other halo-effects of social desirability. Even the definition of membership or non-membership in a predefined group can be sometimes subject to bias. As long as the assessment of the criteria is methodologically deficient, it should not be astonishing to find low correlations between the proposed measure and its intended criteria. By refining the criteria to make these more predictable and convincing as well as improving the proposed measure, the test-criterion correlation can be further enhanced. We still have not reached a sophistication where this multiple-approach of validation by correlating the scores on the proposed test with a set of external criteria is too coarse or obsolete (cf. Thurstone, 1955, p. 356).

The multiple-criteria approach using the method of "known-groups", for example, can be useful not only in validating proposed measures of specific abilities, interests and values but also measures of certain specific attitudes that are generally conceived as behavioural intentions (Scott, W.A. 1968, p.252). For example, in validating a test of achievement value, some of the psychological characteristics of the achievement-oriented person (Mukherjee, B.N.1969) can be used as multiple criteria. However, it would be misleading to infer the extent of validity from the size of the resulting 't' or 'F' statistics used for testing the significance of difference between the groups or from the magnitude of the point-biserial correlations because these statistics depend on the size of the samples and on how the samples are chosen. The research worker must also take care to eliminate the sources of contamination discussed at length by Bellows, R.M. (1941). With adequate criterion-research, including new thinking about criteria at the theory level, and development of more stable as also intrinsically more valid criteria, it would be possible for us in not only using item-validity index for each item but also a better understanding of items by using what Conrad, H.S. (1948, p.48) called 'cross item analysis'. This is done by analysis of an item against more than one criterion. The results emerging from the comparison of item-analysis against an internal consistency criterion and an external criterion may be also quite helpful for item classification. In any event, such type of studies would stimulate further efforts toward individual item improvement and refinement of criteria. As the research takes solid feet, there will be opportunities for a new kind of educational assessment as well. In the area of criterion-referenced testing (CRT), such studies are also likely to help in defining the concept of 'mastery' and in tackling the issue of an *a priori* standard. These two are important but closely interwined problems in CRT.

Educational Evaluation and Examination

Research in educational evaluation and examination has been previously reviewed by Dave, P.N. (1968); Buch, M.B. (1972); Buch, M.B. and Passi, B.K. (1974); Passi, B.K. and Padma, M.S. (1974); Natarajan V. and Kulshrestha, S.P. (1983); Passi, B.K. and Hooda (1986) and more recently by Singh, P and

Prakash, V. (1991). These surveys reviewed the work in these areas in India up to 1988. The present report takes into consideration the Indian studies done during 1988-92. During this period, only six studies on educational evaluation and ten studies directly related to examination could be located. Studies in the areas of prediction of educational achievement, diagnostic tests, selection and admission, as well as promotional studies are not included here.

Studies in Educational Evaluation

Chitnis, S. and Velaskar, P. (1988) examined the qualitative aspects of the educational situation in Maharashtra. According to them, despite the quantitative advances in education, Maharashtra presently harbours serious regional, gender and caste imbalances. There is an all-round erosion in the quality and standard of education. The authors have also offered many suggestions, including the need for flexible syllabi and curriculum at all levels. However, these suggestions are not data-based. Much of the conclusions are based on the authors' intuition and experience.

Veerkar, P. (1989) evaluated the B.Ed. colleges in Maharashtra State by means of a questionnaire which included 75 questions. A stanine scale was prepared for rating the colleges on the basis of scores obtained on the questionnaire and the stanine scale was converted into a 5-point scale. While the scaling of the evaluative adjectives seems to be satisfactory, the reliability and validity of this scale are not ascertained.

Rao, R.S. and Bharathi, M. (1989) made an evaluation of the continuous-evaluation system (CES) of examination in Kendriya Vidyalayas by studying the effect of such continuous evaluations on the final performance of the students. For this purpose, the marks obtained by the students in the ninth class (in five components of CES in five subjects) studying in three different schools, one each from Sambalpur, Visakhapatnam and New Delhi (JNU Campus), were collected from the school records. The marks of these students in these subjects at the Class X Central School Board Examination were also obtained. Although the authors used such evaluative terms as 'nominal success', 'partial success' and 'complete failure' of the CES, no formal scaling method was used for this purpose. There was also no attempt to derive an independent criterion of success which could be used for the purpose of validating the conclusions. The study suffers from the fact that the three schools were not matched in terms of 'benchmark' data, regional differences, and socio-economic variables.

More or less the same methodological pitfalls are evident in the evaluation of the non-detention system (NDS) conducted by Nirmala, J.M. (1989). Without matching the two groups of students of Class X, one under the detention system (N=2,714) and the other under the non-detention system (N=5,026), in terms of achievement of students, their attitude toward the non-detention system, drop-out rates before the introduction of the NDS, the author made a comparative study of the two groups and came to the conclusion that the achievement of students was better under the NDS than under the detention system. Such a conclusion is unwarranted and the evaluation, in this case, is not serving the purpose of a diagnostic tool.

Reddy, Venkata Rami A. and Naidu, G.B. (1988) had earlier published their report of an evaluation of the NDS in the July 1988 issue of *Indian Educational Review*, based on a sample of 2,808 students of Class X of Chittor District drawn by stratified simple random sampling. In this evaluation also, no matching data were collected. The reliability of the marks obtained by the students of the two systems was also not assessed. Because of these methodological deficiencies, it is not possible to ascertain the superiority of one system over the other.

In both the above studies, the goals of the NDS were only implicitly stated. For any evaluation, it is desirable that the specific objectives and goals are stated explicitly before the programme (here, the NDS) commences. Failure to establish clear, specific and measurable objectives is the greatest weakness of many Indian intervention programmes in the field of education. If the programme objectives are stated vaguely or they are non-specific, it becomes difficult to evaluate the programme. This type of deficiency is more glaring in the case of Shah, J.H. and Patel, Y.'s (1989) evaluations of the B.Ed. vacation course. Before the course started, it was necessary for the evaluators to specify the stated goals of the programme and make a detailed study of the expectations, aspirations and state of readiness on the part of the trainees (here, untrained postgraduate teachers working in higher secondary schools). Mere ratings of dissatisfaction about the course at the end of the first winter vacation were not enough, especially because of the possibility of their being faked, random and sensitive to various situational factors. The sampling, moreover, was purposive in this study. The percentage of non-response also was quite substantial. In addition to satisfaction/ dissatisfaction about the vacation course, the achievement of the participants should also have been taken into consideration. Because of these methodological pitfalls, it seems doubtful if Shah, J.H. and Patel, Y. (1989) could really assess the worth of the vacation course.

Kohler, E. (1991) made use of observations, interview and document analysis for the purpose of gauging the standard of occupational therapy (OT) courses in India. She also used a written tool combining the official World Federation OT standards and a self-study checklist for quantitative and qualitative investigation of programme practices and discrepancies in the area of OT. The use of multiple approaches is praiseworthy. The study could have been much more useful if the author had made use of Bloom, B.'s (1956) taxonomy of educational objectives as one of the frames of reference.

Methodological Pitfalls in Educational Evaluation

As in the case of educational assessment, the field of educational evaluation in India is still in its infancy. There is still no clear recognition on the part of many educational researchers in India that some strong and immediate steps must be taken to improve the methodology of educational evaluation if it is to have a respectable future. Most educational evaluations have failed to tackle the problems of conceptualisation, measurement, sampling, and the knotty question of generalisability. In addition to conceptual confusions regarding different aspects of assessing the effectiveness, efficiency and appropriateness of any curriculum, programme and different types of learning outcomes, research in educational evaluation is replete with instances where the principles of evaluation design are flouted. Only in a few evaluations, the benchmark data are collected and the objectives of the programme are spelt out in terms of concrete learning outcomes specifying the behaviour we are willing to accept as evidence of the desired learning. Seldom has there been a set of rigorous experiments under controlled or even quasi-control settings in which children, a classroom or a group of schools were randomly assigned to one or two or more methods of improving education in order to obtain an unbiased estimate of the programme effects and its cost, as has been recommended by Boruch, R.F. (1975). Moreover, too little attention has been paid so far to the importance of carry-over effects in evaluations involving the repeated use of the same subjects. Although the systematic longitudinal method is the ideal procedure of educational evaluation, very few evaluations in India have used this method.

Not a single study could be located in Indian educational research literature which could suggest a comprehensive conceptual framework for educational evaluation in India for the benefit of students, teachers, school administrators, parents and policy makers. The literature shows a dearth of articles highlighting theoretical reflections. Very few empirical studies related to issues involved in the evaluation of educational progress, educational institutions and educational personnel could be located in the literature. There is also no single report summarising the state-of-the-art concerning specific topics in evaluation in general, or, in particular, educational programmes, as in the case of evaluation of social action programmes (Mukherjee, B.N. 1979). The available reports are all fragmentary in nature. This is true even of the recent Report of the Expert Group (1994) evaluating the literacy campaigns in India set up by the Ministry of Human Resource Development, Government of India, New Delhi. The group, no doubt, identified some of the "serious weaknesses" of the campaigns in terms of misreporting, lack of creditability of the monitoring system and pointed out considerable drop-out and relapse to illiteracy on the part of programme beneficiaries but failed to analyse why these negative trends are reappearing (Banerjee, S. 1994, p. 3016). Evaluation reports should be diagnostic, analytic and must go beyond a mere description of the programme strategy, beyond judgment about its effectiveness, to decisions about the necessity of choosing alternative courses of action or the future plan of action in order to bring about a significant improvement in human resource development, in general, and the teaching-learning process in particular. Unless the evaluation report is action-oriented, it is meaningless. Only when it is written in such a readable form that it could be appreciated even by intelligent laymen and used by administrators for improving any action programme, that it becomes a valuable tool for management.

Some Trends in Educational Evaluation in India

The unfortunate trend of treating evaluation as a simple procedure of assessing the worth of an instructional programme is still continuing in India. Very few educational researchers in India fully appreciate the fact that evaluation is not as simple and straightforward as some people seem to believe. Drawing definitive conclusions from studies which compare various instructional methods is quite difficult even when great care is taken to plan the study in advance and use a seemingly appropriate study design (Gage, N.L.1963; Campbell, D.T. and Stanley, J.C. 1963). Drawing inferences from a fortuitous collection of data based on purposive sampling, and non-use of a control group, and without defining in advance precisely what is to be measured as evidence of the programme outcome is almost always not a methodologically sound procedure. Such evaluations not only yield misleading results but may also be dangerous.

Another trend which is clearly noticeable in many educational evaluations conducted in India is that these studies are primarily concerned with the product of learning rather than with the process of learning. The scope of educational evaluation extends much wider than simply determining the correspondence between academic performance and curricular objectives. In revising a curriculum or an instructional strategy, it is very important to know why certain learning outcomes did, or did not, occur, as in the case of relapse to illiteracy. Such type of a process evaluation based on observations and description of how a programme is being conducted and how the change, if any, is occurring in the teaching-learning process is as important as the impact studies where a terminal assessment is made mainly in terms of the programme outcome.

The trend of neglecting the study of negative and/or unanticipated consequences or outcomes of any educational intervention programme is also continuing in India. Evaluation reports discussing unexpected or negative benefits or fortuitous aspects of a programme's operation in a particular type of educational institution can be really revealing. Such discussions may also help in planning and modifying intervention programmes in other systems. In any case, the reporting of unexpected benefits has potential for wider applications elsewhere.

Educational researchers in India, by and large, are still using the traditional system of classifying evaluation variables as dependent and independent. The organismic variables are also being considered now. Mitzel, H.E. (1960) has identified three kinds of evaluation variables: (1) product variables which emphasise the quality of student progress or improvement in the learning outcome, (2) process variables, those which have to do with the actual process of training or intervention, including use of resources, quality and quantity of interactions; and lastly (3) the presage variables, those which are connected with the quality of faculty and students, historical and environmental elements and resources available internally and externally. Even the conceptual framework suggested by the National Institute of Adult Education (1992) fails to make a mention of these variables, although it emphasises the need for a holistic coverage of all aspects of the national total literacy campaigns of the Government of India instead of the prevailing system of "attempting only learner evaluation, cost-effectiveness evaluation and other similar task-related forms of evaluation". Unless the evaluation provides the policy-makers and/or administrators the knowledge relevant to the control of the physical environment of the programme, knowledge pertaining to the programme operation and its beneficiaries and knowledge relevant to the control themselves in terms of their directions and focus (Dror, Y. 1971), the evaluation is just an eyewash or, at best, an academic exercise. The evaluation should identify the ways in which outcome measurements can be used to carry out several policy and management purposes.

Studies Connected with Examinations

Two studies were reported during the period under review which pertain to malpractices in examinations. The first one is by Natarajan, V and Arora, A. (1989) while the second one is by Choudhari, V.M. (1990). While Natarajan, V. and Arora, A. (1989) tried to ascertain the magnitude of unfair means in Indian university examinations, Choudhari, V.M. (1990) confined his study to the relative incidence of malpractices in different faculties of the Nagpur University during 1984-88. Although Natarajan, V. and Arora, A. (1989) attempted to identify some of the factors underlying the use of unfair means in examinations, such as excessive competition and psychological frustrations, by interviewing vice-chancellors, registrars and controllers of examinations, they could hardly pinpoint the possibility of more specific causes at the individual examinee level such as inadequate facilities at home, parental supervision, peer group, influence, study-habits, etc. Such types of investigative study would have been more useful. The data gathered from 35 universities were also not thoroughly analysed. Choudhari, V.M. (1990) also failed to present a trend analysis of the data to show how the incidence of malpractice was changing over time.

During the period under review, two articles were published, which are in some way related to the consistency of examination marks. Patel, N.P.; Patel, S.R. and Patel, R.J. (1990) in their paper entitled "On statistical analysis of examination pattern" reported the results of a study showing the difference in the proportion of students passing three examination systems at the Gujarat Agricultural University, namely, the annual system, the 4-point system and the 10-point system during the academic years 1973 to 1987. Among these three systems, the proportion of students getting first, second and third classes under the 10-point system was higher than under other systems in all the three campuses of the university. However, this result does not necessarily imply that among all the three systems, the 10-point system was the best, as has been actually concluded by the author. The results of the study are also not properly interpreted. The application of the conventional t-test, F-test and chi-square test is also questioned since the assumption of independence of the same samples under the three systems is violated.

Patel, R.S. (1989) studied the degree of disparity between the results of the Class X and Class XII examinations, within the same examinees, from the records obtained from the Gujarat State Examination Board for three schools of Ahmedabad. Without using any multivariate procedure of data analysis, the author came to the conclusion that English, biology, physics and mathematics are the specific subjects (papers) which cause lowering of the percentage of marks at the higher secondary level as compared to the ones secured at the Class X level. This conclusion is technically unsound since all these papers are moderately intercorrelated and yet their intercorrelations are not of the same magnitude. When multiple measurements are taken on the same respondents and/or when they are exposed to multiple overlapping treatments, these measures are often found to be moderately correlated. The classical univariate treatment of such data separately for each variable is not justified because a set of correlated data, in general, leads to departure from the assumptions of statistical independence and additivity of treatment effects.

Thakre, V. (1991) made a comparative study of revaluation of 2,29,173 answer books obtained from various faculties of Nagpur University and 1,638 answer books from Punjabrao Krishi Vidyapeeth, Akola, during 1985-90. Without making any time series analysis or running any test of trend, she came to the conclusion that every year the cases of revaluation are on the increase in both the universities. She also did not make a detailed analysis of the type of mistakes made by the valuers (examiners).

Sujatha, B.N. (1991) also made a study of revaluation of answer books of undergraduate engineering students who appeared in various semester examinations during 1982-88. She also did not make an intensive analysis of the data and, without proper controls, came to the conclusion that revaluation does have an effect on the marks and results.

Barua, A.P. (1989) made an opinion survey of the causes underlying failure in the HSLC examination in Assam without using a proper sampling design of different types of respondents. Since only 35% of the sampled respondents returned the filled-in booklets, it certainly is not a representative sample. Although the study addressed an important problem, it failed to make a proper contribution because of methodological deficiencies. Also, the study was not guided by any relevant theoretical framework connected with the phenomenon of large scale failures in public examinations.

Modi, D.J. (1989) made an attempt to set up an item bank in Gujarati language for Class X students studying in Bhavnagar City. The try-out form of the test was administered to 3,340 students of 22 schools, and facility value and discrimination index were calculated for objective and short-answer type questions. Distracter analysis was also done. The final pool of items consisted of 50 essay-type questions, 110 short-answer type questions and 256 objective-type questions, covering prose, poetry and grammar. However, there was no attempt to develop a few parallel form tests out of selected questions. There was also no attempt to study the time which should have been allotted for attempting each question.

Malhotra, M.M., Bedi, S.P. and Tulsi, P.K. (1990) made a content analysis of question papers set in the Board Examination of Haryana polytechnics in order to asses the coverage of the syllabus as well as the adequacy of the weights assigned to questions aimed at measuring different levels of cognitive abilities. However, these authors used only a few experts for this purpose. They also did not use any procedure for making a thorough breakdown of the content areas of the syllabus and of the outcomes of learning these content areas in the form of a table of specifications. Similarly, there was no rational procedure adopted for arriving at the weights with respect to comprehension, application and other outcomes. For this reason, the content validity of each question paper could not be established. The difficulty value and discriminatory power of each question were also not obtained since this was beyond the scope of the study. Notwithstanding these, the attempt on the part of the authors to study the sequencing of the questions and the adequacy of the instructions for the different parts, and to identify the ambiguity involved in each is quite commendable.

Malhotra, M.M., Menon, P.N., Bedi, S.P. and Tulsi, P.K. (1989) made a status study of the existing system of internal assessment of students in the polytechnics of Haryana. For this purpose, they sampled 35 teachers and 145 students of three polytechnics and using an interview schedule and a questionnaire (for students), the authors attempted to measure their level of knowledge of the criteria used for evaluating their performance and their attitude toward the different aspects of practices connected with internal assessment. Although no attempt was made to establish the reliability and validity of the tools employed in the study, the findings of the study appear to be quite useful.

Using the experience of the above-mentioned status study, Malhotra, M.M. and Tulsi, P.K. (1990) attempted to design an improved system of internal assessment of students in the polytechnics of Haryana. The system recommended by them included a scheme for monitoring students' progress, and the use of feedback of assessment results to students, teachers, curriculum planners and administrators. Steps for implementing the scheme are also suggested in the design. This follow-up study is also to be commended for providing a guideline for designing tools for the various components of internal assessment.

During the period under review, two studies dealing with assessment of attitudes toward oral examination and toward spot evaluation were reported. While Thangamani, C.C. (1989) studied the attitude of higher secondary school teachers of Madurai toward oral examination *vis-a-vis* written examination, Rai, V.K. (1989) analysed the attitude of examiners towards spot evaluation in relation to sex, level and stream. In both studies, a Likert-type attitude scale was developed but the formal psychometric procedure was not described. In both cases, the samples used were of small size. The sampling also appears to be purposive in both studies. The problem of response-set toward either agreement or disagreement with whatever statement was made was also not taken into consideration in both these studies. Both studies are, therefore, methodologically deficient.

Kumar, S. (1991) made a comparative study of grading and marking with respect to the consistency of results where students' performance is evaluated by general examiners. For this purpose, a purposive sample of 30 teacher trainees was selected and 55 answer scripts were examined by them for marking, re-marking, grading and regrading. Although Kumar, S. (1991) found marking to be as reliable as grading, the reliability in the two cases being 0.74 and 0.77 respectively, he failed to examine the extent of inter-examiner and intra-examiner variability for the two procedures of assessment. The means and their standard errors are also not reported for grading, regrading, marking and re-marking. Whether the reported reliabilities are based on the Fisher-Z transformation of correlations between grading (marking) and regrading (re-marking) of examinees by the same examiner pooled across the sample of 30 examiners is also not clear. For this reason, it is difficult to accept the conclusion of Kumar, S. (1991) claiming that the conventional marking system to be as reliable as the grades based on relative ranking of examinees. Previous studies have shown grades on rank order to be more reliable than the conventional marks and, for all practical purposes, the joint use of both grades and marks has been found to be quite informative and useful.

However, there should be more studies in the area of grading, scaling of marks, error of decision-making with regard to each, and how these are influenced by types of examiners, subject-matter, types of questions, etc. Studies on the variations from board to board in conducting different public examinations and a thorough re-analysis of their year-wise results should be undertaken. Follow-up studies should be taken up to validate some of the conclusions arrived at by the previous investigations undertaken by the Directorate of Extension Programmes of Secondary Education (DEPSE). The feasibility of using the computer for generating questions from item-banks as well as for scoring and storing the examination marks in a strict environment of confidentiality should be studied. Efforts should also be directed towards increasing the reliability and validity of essay-type questions.

CONCLUDING REMARKS

By and large, the quality of assessment procedures, which are currently being used in India for educational assessment and evaluation, is far from being technically sound and/or innovative. The standard set so far has largely been discouraging and leaves much to be desired. Despite this, it is surprising that there has not been much resentment or reaction in India against psychological and educational testing. Even the general public has not been much concerned over the way important decisions are being made year after year about the future careers of our students and teachers based on shaky devices. In other countries, the use of educational assessment and evaluation procedures has evoked an increasing flow of resentment and critical comments. In India, the negative as well as the positive consequences of educational testing have seldom been debated. Educational researchers, with few exceptions, are still depending more on tests rather than on other methods of assessment. The only encouraging trend in this respect is the growing appreciation of the CRT *vis-a-vis* normative testing.

The use of the computer in educational research for other than routine tasks has been minimal in India. In the developed countries, the computer is increasingly being used to assist the individual student by estimating the extent of his/her resemblance with a known group, by warning of inconsistencies between the expectation and the performance of the student, by recommending appropriate learning experiences, by assisting in his placement and by identifying skills or concepts requiring remedial attention or relearning (Cooley, W.W. 1964). In addition to these guidance functions, computers are being used in the developed countries for the purpose of administration of tests, generation of items and computer-assisted test assembly. The use of video-tape closed-circuit TV, telemetry instruments and other recent technological advances in instructional technology has not so far brought notable progress in India in the field of educational assessment, examination and evaluation.

There has also been little interest in the development of new theoretical approaches to educational assessment such as; Cronbach, L.J. and Gleser G.'s (1966) decision-theoretical approach; Jensen A.R.'s (1970) two-level theory of ability which according to Cronbach, L.J. (1975) has generated unparallel controversy in the history of mental testing; Das, J.P. et al.'s (1979) theory of simultaneous and successive cognitive processes; Gough, H.G.'s (1965) three-stage test evaluation process, adaptive testing, etc.

Against the background of the current progress in matrix sampling (Shoemaker, D.N. 1975), statistical modelling in school effectiveness studies (Aitkin, M. and Longford, N. (1986) and computerised adaptive testing (Bejar, I.J. 1985), the technique of research connected with educational assessment in India is narrowly conceived and largely obsolete. In most examples of educational testing in India, even now little emphasis is being placed either on testing or on education.

The state-of-the-art so far as the methodological aspects of assessment and evaluation are concerned is still almost primitive. Virtually no systematic attempt has been made so far to develop innovative procedures, such as unobtrusive measures (Webb, E.J., Campbell, D.T., Schwartz, R.D. and Sechrest, L. 1966), or the psycho-physiological assessment of stress as also dispositional variables during the teaching-learning process and in creativity. Except in a few cases, no systematic attempt has been made to develop adequate measures of the learning environment. One hardly notices even a shift from the use of individual measures of abilities, interest, achievement, motivation, etc., to measures of learning environment which in recent years are being recognised as potential predictors of learning outcomes.

We still have a long way to go before the recent spectacular advances in computer and scoring technology can be fruitfully exploited in a routine manner by our universities and other organisations conducting public examinations. Only in a few institutions, the computerised scoring technology presently in vogue is bringing about some anticipated change in item construction and test analysis. There is not much recognition of the suggestion that future efforts be directed towards individual item improvement, domain sampling and scaling procedures, especially multi-dimensional scaling.

So far as the conceptual and pedagogic aspects connected with any educational assessment are concerned, there have been only a few instances where an attempt has been made to penetrate deeply into the complexities of the problem. There is, almost no systematic attempt to handle the conceptual problem involved, for example, in the assessment of creativity and to integrate the existing theories such as the one proposed by Jackson, P.W. and Messick, S. (1965). The feedback from the existing pedagogical and psychological theories in shaping assessment procedures has been of limited value. The results obtained from uncritical methods of data gathering in evaluation studies without any theoretical backing are likely to mislead if used for policy decisions. In any event, such theoretically and empirically lopsided studies are responsible for educational research in India making very slow progress. The organisation and integration of research that bears upon the assessment of important constructs of education (such as learning outcome, teacher effectiveness, creativity, learning environment, etc.) have also been inhibited by the absence and/or minimal utilisation of a relevant theoretical framework. For this reason, educational researchers in India still have some uncertainty about: (a) what type of variables they should include in their investigations, (b) how they can best classify responses connected with learning outcomes and/or different areas of their interest, and (c) how to interpret these responses and study the value of such data in making wise decisions in the face of uncertainty. To some extent, this neglect in not using a relevant theoretical framework has not only hampered the flow of cumulative research in India but has also created uncertainty in the minds of many educational researchers as to what constitutes adequate research in the area of test validation. This, in turn, has resulted in a large number of irrelevant and poorly designed studies in education that call for a total ban on indiscriminate publishing. The sheer bulk of publications could indeed be reduced if the papers were scrutinised by a small group of hard-nosed methodology specialists.

In view of what has been discussed above, this reviewer is compelled to conclude that the overall state-of-the-art in educational assessment and evaluation in India is far below the level of sophistication necessary for making educational testing develop into an emerging science of assessment. Educational tests, tools and evaluation have still to emerge in India as a set of instruments for policy research. It will take a long time before assessment procedures are woven into the fabric of education. It still is not functioning even as a means to facilitate our understanding of the pedagogical and psychological processes involved in any teaching-learning process. The present examination system and procedures of internal and external evaluation are also so narrow that the entire teaching and learning process is geared to passing examinations and getting good marks required for entry to the higher level of education and/or the job market. They are hardly serving the purpose of conveying powerful messages to teachers, students and parents about what should be taught. Such type of feedback could have provided our policy- makers with information as to how educational assessment can play the role of a cost-effective tool for improving the quality of education.

In order to fulfil the expectation that properly validated and standardised educational tests (by virtue of their objectivity and freedom from some of the common biases) can substantially improve the human resource development programme in India, we need to strengthen our research methodology training efforts, as has been suggested elsewhere (Mukherjee, B.N. 1992). Besides improving the educational researcher's understanding of the basic principles of the scientific method, such training programmes should aim at presenting systems and approaches (including the functional use of the computer) that can be used not only to plan technically sound research projects in education but also to help in assessing the validity of research findings and undertake innovative studies in the area of indigenous assessment methods.

REFERENCES

Ambasana, Anil Dhirajlal. 1989. **Construction and standardisation of an art judgement test.** Ph.D., Edu. *Saurashtra Univ.*

Aitkin, M. and Longfcrd, N. 1986. **Statistical modelling issues in school effectiveness studies.** *Journal of the Royal Statistical Society*, Series A, Vol. 149, 1-43.

Ashai, Yasmeen and Mohite, P. 1989. **Establishing norms for the teacher's rating scale.** *Indian Educational Review*, Vol. 24(4), 60-67.

Bakan, D. 1966. **The test of significance in psychological research**. *Psychological Bulletin*, Vol. 66, 423-437.

Banerjee, S. 1994. **Flowers for the illiterate**. *Economic and Political Weekly*, Nov. 26, 3013-3016.

Barua, A.P. 1989. **Causes of failure in Higher Secondary Leaving Certificate Examination.** Independent study. *Assam.: State Council of Educational Research and Training.*

Basumallik, T.; Bhattacharya, K.P.; Banerjee, S. and Mitra, S.K. 1992. ***Assessment of minimum learning in primary education***. Calcutta: Indian Statistical Institute.

Behera, N. 1990. **Standardization of adjustment projective-inventory.** M.Phil., Psy. *Utkal Univ.*

Bejar, I.J. 1985. **Speculations on the future of test design**. In, S. Whitley (Ed). ***Test Design ; Contributions from Psychology, Education and Psychometrics,*** New York: Academic Press.

Bellows, R.M. 1941. **Procedure for evaluating vocational criteria**. *Journal of Applied Psychology*, Vol. 25, 499-513.

Bloom, B.S. (Ed.) 1956. ***Taxonomy of Educational Objectives***. New York: Longmans, Green & Co.

Bock, R.D.; Dickens, C.; and Van Pelt, J. 1969.

Methodological implications of content-acquiesence correlation in the MMPI. *Psychological Bulletin*, Vol. 71, 127-139.

Boruch, R.F. 1975. **Coupling randomized experiments and approximations to experiments.** *Sociological Methods and Research*, Vol. 4, 31-53.

Boruch, R.F. and Wolin,L. 1970. **A procedure for estimation of trait, method and error variance attributable to a measure**. *Educational and Psychological Measurement*, Vol. 30, 547-574.

Bramble, W.J. and Wiley, D.E. 1974. **Estimating content acquisence correlation by covariance structure analysis**. *Multivariate Behavioural Research*, Vol. 9, 179-190.

Buch, M.B. 1972. ***Educational Psychology. A Survey of Research in Psychology*** Bombay: Popular Prakashan, 80-125.

Campbell, D.T. and Fiske, D.W. 1959. **Convergent and discriminant validation the multi-trait-multimethod matrix.** *Psychological Bulletin*, Vol. 56, 81-105.

Campbell, D.T. and Stanley, J.C. 1963. **Experimental and quasi-experimental designs for research on teaching.** In, N.L. Gage (Ed.) ***Handbook of Research on Teaching***. Chicago: Rand McNally, 171-246.

Canter, D. 1985. (Ed.) ***Facet Theory: Approaches to Social Research.*** New York: Springer-Verlag.

Chauhan, Radha. 1988. **Construction and standardization of an academic alienation scale and its relationship to demographic variables.** Ph.D., Edu. *Himachal Pradesh Univ.*

Chawla, Swarn. 1988. **Construction of a Multiple-choice Reading Comprehension Test.** *Indian Educational Review*, Vol 23(4), 1-17.

Chawla, Swarn. 1992. **Standardisation of a Multiple-choice Vocabulary Test: Pre-test and analysis.** *Indian Educational Review*, Vol. 27(3), 49-68.

Chitnis, S. and Velaskar, P. 1988. **Education in Maharashtra: Strengths and weaknesses.** Independent study. *Bombay: Tata Institute of Social Sciences.*

Choudhari, V.M. 1990. **A comparative study of malpractices in examinations during 1984 to 1988 in Nagpur University.** M.Phil., Edu. *Nagpur Univ.*

Conrad, H.S. 1948. **Characteristics and uses of item analysis data**. *Psychological Monograph*, No. 295, Washington: American Psychological Association.

Cooley, W.W. 1964. **A computer-measurement system for guidance**. *Harvard Educational Review*, Vol. 34, 557-572.

Cooil, B. and Rust, R.T. 1994. **Reliability and expected loss**. *Psychometrica*, Vol. 59, 203-216.

Cronbach, L.J. 1957. **The two disciplines of scientific psychology**. *American Psychologist*, Vol. 12, 671-684.

Cronbach, L.J. 1975. **Five-decades of controversy over mental testing**. *American Psychologist*, Vol. 30, 1-14.

Cronbach, L.J. and Gleser, G. 1966. ***Psychological Tests and Personnel Decisions***, 2nd Edn., Urbana, Ill: University of Illinois Press.

Cronbach, L.J.; Gleser, G.; Nanda, H. and Rajaratnam, N. 1972. ***The Dependability of Behavioural Measurements: Theory of Generalizeability for Scores and Profiles***. New York: Wiley.

Cruise, Robert J. 1988. **Research consideration in setting effect size.**

Indian Educational Review, Vol. 23(1), 1-6.

Das, J.P.; Kirby, J.R. and Jarman, R.F. 1979. ***Simultaneous and Successive Cognitive Processes.*** New York: Academic Press.

Dash, U.N. and Das, J.P. 1984. **Development of concrete operational thought and information coding in schooled and unschooled children**. *British Journal of Developmental Psychology*, Vol. 2, 63-72.

Dave, Meeta. 1992. **An investigation into reading comprehension of the pupils of Class VII by using the standardised tests in Gujarati.** Ph.D., Edu. *Gujarat Univ.*

Dave, P.N. 1968. **Educational evaluation and examination: A trend report**. In, M.B. Buch (Ed.) ***Fourth Survey of Research in Education***. New Delhi: National Council of Educational Research and Training.

Dawyer, J.H. 1974. **Analysis of variance and the magnitude of effects: A general approach**. *Psychological Bulletin*, Vol. 81, 731-737.

Dodd, D.H. and Schultz, R.F. 1973. **Computational procedures for estimating magnitude of effect for some analysis of variance designs**. *Psychological Bulletin*, Vol. 79, 391-395.

Dror, Y. 1971. ***Ventures in Policy Sciences.*** New York: American Elsevier.

Ebel, R.L. 1961. **Must all tests be valid?** *American Psychologist*, Vol. 16, 640-647.

Foa, U.G. 1965. **New developments in facet design and analysis**. *Psychological Review*, Vol. 72, 262-274.

Foa, U.G. and Turner, J.L. 1970. **Psychology in the year 2000: Going structural**. *American Psychologist*, Vol. 25, 244-247.

Gage, N.L. 1963. **Paradigms for research on teaching.** In, N.L. Gage (Ed.) ***Handbook of Research on Teaching***. Chicago: Rand McNally.

Glass, G.V. and Hakstian, A.R. 1969. **Measures of association in comparative experiments: Their development and interpretation**. *American Educational Research Journal*, Vol. 6, 403-414.

Gough, H.G. 1965. **Conceptual analysis of psychological test scores and other diagnostic variables**. *Journal of Abnormal and Social Psychology*, Vol. 70, 294-302.

Guilford, J.P. 1965. **The structue of intellect**. *Psychological Bulletin*, Vol. 53, 267-293.

Guilford, J.P. 1988. **Some changes in the structure of intelligence model.** *Educational and Psychological Measurement*, Vol. 48.

Roy Guha, S.; Mitra, Subir K., and Ray, S.S. 1995. ***Achievement Level of Primary School children at the end of Class IV.*** Calcutta: Indian Statistical Institute & State Council of Educational Research and Training, West Bengal.

Guttman, L. 1954. **A new approach to factor analysis: The radix.** In P.F. Lazarsfeld (Ed.) ***Mathematical Thinking in the Social Sciences***. Illinois: Free Press, 258-348.

Guttman, L. 1965. **A faceted definition of intelligence**. In, R. Eiferman (Ed.) ***Studies in Psycholgy: Scripta Hierosolymitana.*** Vol 14, Jerusalem, Israel: The Hebrew Univ.

Guttman, L. 1971. **Measurement as structural theory**. *Psychometrika*, Vol. 36. 329-348.

Guttman, R. and Schlesinger, I.M. 1967. **The analysis of diagnostic effectiveness of a facet design battery of achievement and analytical ability tests.** *Jerusalem, Israel: The Israel Institute of Social Research.*

Harper, A.E. Jr. 1960. ***Recent advances in psychometry***. Silver Jubilee Volume of the Journal of Vidya Bhavan Society, Udaipur.

Harris, T.L. 1962. **Some issues in beginning**

reading instruction. *Journal of Educational Research*, Vol. 56, 5-19.

Jackson, D.N. and Messick, S. 1968. **The logic of assessment**. In, D.N. Jackson and S. Messick (Eds.) ***Problems in Human Assessment***, New York: McGraw-hill, 41-42.

Jackson, P.W. and Messick, S. 1965. **The person, the product, and the response: Conceptual problems in the assessment of creativity**. *Journal of Personality*, Vol. 33, 309-329.

Jain, S.C. 1989. **Piagetian theory of intellectual development and format of measurement**. A paper presented at a National Seminar on *The issues and problems of Psychological Testing*, held in 1992 at National Council of Educational Research and Training: New Delhi.

Jalota, S. 1965. **Intelligence testing in India**. *Indian Psychological Review*, Vol.1. 96-107.

Jensen, A.R. 1970. **Hierarchical theories of mental ability**. In, Dockrell, W.B. (Ed.) ***On Intelligence: The Toronto Symposium, 1969***. London: Methun.

Jyoti, Nirmala, M. 1992. **An evaluation of the non-detention system.** Ph.D., Edu. *Sri Venkateswara Univ.*

Kiran, G. and Lewis, C. 1979, **Partial omega squared for ANOVA designs.** *Educational and Psychological Measurement*, Vol. 39, 119-128.

Khan, Yusuf. 1989. **Construction and standardisation of diagnostic tests in English for Standard VIII with regard to structures.** Ph.D., Edu. *Nagpur Univ.*

Khire, Usha. 1989. **Construction of a battery of tests based on Guilford's S.I. model.** Independent study. *Pune: Jnana Prabodhini Institute of Psychology.*

Kohler, Elizabeth. 1991. **Occupational therapy—educational standards in India: A case and field study.** *Indian Educational Review*, Vol. 26(2), 1-9.

Kulkar, K.R. 1989. **Construction and standardisation of unit tests in the subject of Marathi for pupils of Standard VIII.** Ph.D., Edu. *Shivaji Univ.*

Kulkarni, S.S. and Kumar, K. 1986. **Tests and measurement: A trend report.** In M.B. Buch (Ed.) ***Fourth Survey of Research in Education***. New Delhi: National Council of Educational Research and Training.

Kulkarni, S.S. and Puhan, B.N. 1988. **Psychological assessment: Its present and future trends**. In, J. Pandey (Ed.) ***Psychology in India: The State-of-the Art.*** Vol 1, New Delhi: Sage Publications, 19-91.

Kumar, Anil. 1990. **Construction and standardisation of performance test of general mental ability for illiterate adults in the age group 15-35 years.** Ph.D., Edu. *Kurukshetra Univ.*

Kumar, K. 1991. **Research in tests and measurement : A trend report.** In M.B. Buch (Ed.) ***Fourth Survey of Research in Education***. New Delhi: National Council of Educational Research and Training.

Kumar, S. 1991. **Comparative reliability of grading and marking.** Independent study. *The Maharaja Sayajirao Univ. of Baroda.*

Lodhi, P.H. 1991. **Analysing relationship among multiple data sets for a single sample: A methodological appraisal.** *Indian Educational Review*, Vol 26(4): 118-23.

Malhotra, M.M.; Menon, P.N.; Bedi, S.P. and Tulsi, P.K. 1989. **Status study of internal assessment of students in the polytechnics of Haryana.** Independent study.*Chandigarh: Technical Teachers' Training Institute.*

Malhotra, M.M. and Tulsi, P.K. 1990a. **Scheme for internal assessment of students in the polytechnics of Haryana.** Independent study. *Chandigarh: Technical Teachers' Training Institute.*

Malhotra, M.M.; Bedi, S.P. and Tulsi, P.K. 1990. **Content analysis of question papers set in the Board Examination of Haryana polytechnics.** Independent study. *Chandigarh: Technical Teachers' Training Institute.*

Malhotra, M.M.; Anand, Y.K. and Bedi, S.P. 1990. **Design of the system of student evaluation incorporating multi-point entry and credit system.** Independent study. *Chandigarh: Technical Teachers' Training Institute.*

Manjula, R. 1991. **Construction and standardisation of a reading readiness test in English for pre-school children.** Ph.D., Edu. *Bangalore Univ.*

Marsh, H.W. 1987. **The hierarchical structure of self-concept and the application of hierarchical confirmatory factor analysis.** *Journal of Educational Measurement*, Vol. 24, 17-19.

Menzel, E.W. 1950. ***The Use of New Type Tests in India.*** (4th Edn.). London: Oxford University Press.

Messick, S. 1970. **The criterion problem in the evaluation of instruction: Assessing possible, not just intended outcomes.** In, M.C. Wittorock and D.E. Wiley (Eds.) ***The Education of Instruction: Issues and Problems***. New York: Holt, Rinhart & Winston.

Milgram, N.A.and Helper, M.M. 1961. **The social desirability set in individual and grouped self-ratings.** *Journal of Consulting Psychology*, Vol. 25, 91.

Mitra, S.K. 1961. **Research needs in the development of different types of tests in India.** In, T.K.N. Menon (Ed.) ***Recent Trends in Psychology***, *Calcutta: Orient Longmans.*

Mitra, S.K. 1968. **Review of tests and measurement**. In, S.B. Adaval (Ed.) ***The Third Indian Yearbook of Education***. *New Delhi: National Council of Educational Research and Training.*

Mitra, S.K. 1972. **Methodology and research technology**. In, S.K. Mitra (Ed.) ***A Survey of Research in Psychology***. Bombay: Popular Prakashan. 414-432.

Mitra, S.K. and Kumar, K. 1974. **Tests and measurement: A trend report.** In, M.B. Buch (Ed.) ***A Survey of Research in Education.*** Baroda: The M.S. University of Baroda.

Mitra, S.K. and Kumar, K. 1979. **Tests and measurement: A trend report.** In, M.B. Buch (Ed.) ***Second Survey of Research in Education***. Baroda: Society for Educational Research and Development, Baroda.

Mitzel, H.E. 1960. **Teacher Effectiveness**. In, C.W. Harris (Ed) ***Encyclopaedia of Educational Research***, New York: Macmillan & Co.

Modi, D.J. 1989. **Question bank for Standard X in Gujarati subject.** Independent study. *Bhavnagar Univ.*

Mohan, S.; Pant, D.; Sibia, A. and Sharma, R.K. 1992. **Psycho-educational assessment in India : Present status**. In, S. Mohan and A. Sibia (Eds.) ***A Report of National Workshop on Psycho-Educational Assessment : Identifying Gaps.*** New Delhi: National Council of Educational Research and Training (Mimeographed).

Mukherjee, B.N. 1966. **Derivation of likelihood ratio tests for Guttman quasi-simplex covariance structures**. *Psychometrika*, Vol. 31, 97-123.

Mukherjee, B.N. 1969. **Some characteristics of the achievement-oriented person: Implications for the teacher-learning process**. *International Journal of Educational Science*, Vol. 3, 209-216

Mukherjee, B.N. 1973. **Analysis of covarinace structures and exploratory factor analysis**. *British Journal of Mathematical and Statistical Psychology*, Vol. 26, 125-154.

Mukherjee, B.N. 1974a. **A questionnaire measure of persistence disposition**. *Indian Journal of Psychology*, Vol. 49, 263-278.

Mukherjee, B.N. 1974b. **Towards a conceptualization of the achievement value construct.** In, S.K. Roy and A.S.K. Menon (Eds.) ***Motivational & Organizational Effectiveness***. New Delhi: Sri Ram Centre for Industrial Relations. 43-74.

Mukherjee, B.N. 1976. **Techniques of covariance structural analysis**. *Australian Journal of Statistics*, Vol. 18 (15), 131-150.

Mukherjee, B.N. 1979. **Theory and method**. In, Udai Pareek (Ed.) ***A Survey of Research in Psychology (1971-76) Part I.*** Bombay: Popular Prakashan, 1-135.

Mukherjee, B.N. 1985. **The reliability and validity of the so-called 'hard' and 'soft' data.** *Journal of Social and Economic Studies*, Vol. 2, 135-191.

Mukherjee, B.N. 1992. **Quality control in Indian educational research**. *Journal of Indian Education*, Vol. 18, 22-34.

Mukherjee, B.N. 1993a. **Needed research in psycho-educational assessment in India**. *Psychological Studies*, Vol. 38, 85-100.

Mukherjee, B.N. 1993b. **Toward a rapprochement between the two basic paradigms of educational research.** *Quality* and *Quantity*, Vol. 27, 383-410.

Mukerjee, D.P. 1991. **Testing reading comprehension: A comparative analysis of cloze test and multiple choice test.** *Indian Educational Review*, Vol. 26(1), 44-68.

Murphy, G. 1969. **Psychology in the year 2000**. *American Psychologist*, Vol. 24, 523-530.

Nandi, A. 1976 **Adorno in India: Revisiting the psychology and fascism**. *Indian Journal of Psychology*, Vol. 51, 168-178.

Natarajan, V. 1984. **An application of item response theory to aid discrimination function in achievement testing**. D. Litt., Edu. *Poona Univ.*

Natarajan, V. and Arora, Asha. 1989. **Unfair means in university examinations: A study.** Independent study. *New Delhi: Association of Indian Universities.*

Natarajan, V. and Kulshrestha, S.P. (Eds.) 1983. ***Evaluation Methodology and Examiantion Reform.*** Dehradun: Jugal Kishore.

National Institute of Adult Education. 1992. ***Evolving an evaluation system for the total literacy campaign in Delhi: A conceptual framework***. New Delhi.

Report of Expert Group National Literacy Mission. 1994. Directorate of Adult Education, Ministry of Human Resource Development, Government of India, New Delhi, September 1954.

Nirmala Jyoti, M. 1989. **An evaluation of the non-detention system.** Ph.D., Edu. *Sri Venkateswara Univ.*

Padmini, T. 1980. **Fostering cognitive development in first standard pupils: An experimental study**. Ph.D. Edu. *Univ. of Mysore.*

Panchal, D.H. 1991. **Adaptation and standardisation of the first half (first twelve) sub-scales of British Ability Scales for the Gujarati population of urban area.** Ph.D., Edu. *Gujarat Univ.*

Passi, B.K. and Hooda. 1986. **Educational evaluation and examinations: A trend report.** In, M.B. Buch (Ed.) ***Fourth Survey of Research in Education.*** New Delhi: National Council of Educational Research and Training.

Passi, B.K. and Padma, M.S. 1974. **Educational evaluation and examinations: A trend report**. In, M.B. Buch (Ed.) ***Fourth Survey of Research in Education***. New Delhi: National Council of Educational Research and Training.

Patel, B.V. 1988. **Construction and standardisation of silent reading comprehension tests in Gujarati for pupils of Classes V, VI and VII.** Independent study. *Sardar Patel Univ.*

Patel, N.P.; Patel, S.R. and Patel R.J. 1990. **On statistical analysis of examination pattern.** *Indian Educational Review*, Vol. 25(4), 10-17.

Patel, R.S. 1989. **An investigation into the disparity of results of examinations of Standards X and XII conducted by the Gujarat State Examination Board.** *Madhyamik Shikshan and Parikshan*, Vol. 12(7),

Patel, S.R. 1991. **Adaptation and standardisation of the second half (other twelve) scales of British Ability Scales for the Gujarati population of urban areas.** Ph.D., Edu. *Gujarat Univ.*

Patnaik, S.P. 1990. **Development and standardisation of situational tests for selection of elementary school teachers.** Ph.D., Edu. *Utkal Univ.*

Puhan, B.N. 1982. ***Issues in Psychological Measurement.*** *Agra: National Psychological Corporation.*

Rai, V.K. 1989. **Attitude of examiners towards spot evaluation in relation to sex, level and stream.** *Indian Educational Review*, Vol. 24(4), 125-132.

Raithththa, Bhanumati C. 1989. **Development and validation of Criterion-referenced test on vowel-coalition (*Svarsandhi*) in Sanskrit language.** Ph.D., Edu. *Bhavnagar Univ.*

Rani, Raj. 1993. **Abilities involved in learning chemistry at the secondary stage: A factorial and validation study**. *Indian Educational Review*, Vol. 28, 97-100.

Rao, R.R.S.P. 1991. **Development and application of a scale for measuring attitudes towards the new pattern of education and empirical validation of its psychometric properties.** Ph.D., Edu. *Utkal Univ.*

Rao, R.S. and Bharathi, M. 1989. **Evaluation of continuous evaluation system of examination in Kendriya Vidyalayas.** Independent study. *Sambalpur Univ.*

Raviya, D.L. 1990. **Construction and standardisation of reading comprehension test in the subject of Sanskrit for Class VIII students of Saurashtra region.** Ph.D., Edu. *Saurashtra Univ.*

Reddy, Venkata Rami A. 1989. **An evaluation of the non-detention system from different angles.** Independent study. *Sri Venkateswara Univ.*

Reddy, Venkata Rami A. and Naidu, Bhaskara G. 1988. **Achievement of students under detention and non-detention systems.** *Indian Educational Review*, Vol 23(3), 41-55.

Rozario, L. 1989. **Construction and standardisation of achievement tests in physics, chemistry and biology for Standards VIII and IX for students studying through English medium in suburbs of Bombay with a view to diagnostic analysis and remedial teaching in Standard IX and its appraisal.** Ph.D., Edu. *Univ. of Bombay.*

Rucker, G. 1987. ***Mind Tools***. New York: Wiley.

Sandhu, **T.S. 1981. A factorial study of adolescent thought using Piaget type tests.** Ph.D., Edu. *Rajasthan Univ.*

Scott, W.A. 1968. **Attitude measurement**. In, G. Lindzey and E. Aronson (Eds.) ***The Handbook of Social Psychology***. Vol. 2, 2nd Edition, 204-274.

Shah, J.H. 1989. **Construction and standardisation of self-concept inventory in Gujarati for pupils of Classes IX and X.** *Experiments in Education*, Vol. XVII (2), 31-36.

Shah, J.H. and Patel, Yashomati. 1989. **Evaluation of B.Ed. vacation course by student-teachers.** *Experiments in Education*, Vol XVII (12), 308-312.

Shah, M.A. and Agarwal, Saroj. 1988. **Construction of four-dimensional risk taking test.** *Indian Educational Review*, Vol. 23(2), 1-14.

Sharma, S.K. 1991. **Development of predictive battery of tests for scientific aptitude for the students of Class XI.** Ph.D., Edu. *Univ. of Jammu.*

Shashilatha. 1977, **Prediction of academic achievement**. Ph.D., Psy. *Meerut Univ.*

Shoemaker, D.M. 1975. **Toward a framework for achievement testing**. *Review of Educational Research*, Vol. 45, 127-147.

Siddiqui, Shahjehan. 1988. **Standardisation of a test in creative thinking for Urdu-speaking students at the formal operative stage in Telangana area, Andhra Pradesh State.** Ph.D., Edu. *Osmania Univ.*

Singh, Narendra. 1988. **Construction of a scale of reading-writing skills of pre-primary children.** Ph.D., Edu. *Agra Univ.*

Singh, Pritam. 1988. **Development of criterion referenced tests in environmental studies (science) for the primary stage.** Independent study. *New Delhi: National Council of Educational Research and Training.*

Singh, Pritam and Prakash, V. 1991. **Research in educational evaluation and examination: A trend report.** In M.B. Buch (Ed). ***Fourth Survey of Research in Education 1983-88***. *New Delhi: National Council of Educational Research and Training.*

Sinha, D. 1986. ***Psychology in a Third World Country : The Indian Experience***. New Delhi: Sage.

Sujatha, B.N. 1991. **Revaluation in B.E. degree examinations: An analysis of marks.** Independent study. *Univ. of Mysore.*

Thakre, Veena. 1991. **A comparative study of revaluation of answer books in the Nagpur University and Punjabrao Krishi Vidyapeeth, Akola, during 1985-90.** Ph.D., Edu. *Nagpur Univ.*

Thangamani, C.C. 1989. **Oral examination as**

an instrument of diagnostic evaluation. M.Phil., Edu. *Madurai Kamaraj Univ.*

Thurstone, L.L. 1955. **The criterion problem in personality research**. *Educational and Psychological Measurement*, Vol. 15, 353-361.

Vaghela, Vrajlal K. 1992. **Development and validation of a criterion referenced test in social studies for Standard VII.** Ph.D., Edu. *Bhavnagar Univ.*

Veerkar, P. 1989. **Preparation of a criterion scale for rating the B.Ed. colleges in Maharashtra State.** Independent study. *Pune: Maharashtra State Council of Educational Research and Training.*

Vyas, Sharad G. 1988. **Construction and standardisation of the Hindi language ability test for the entrants to Standard XI (general stream) of higher secondary schools of Saurashtra region.** Ph.D., Edu. *Bhavnagar Univ.*

Webb, E.J.; Campbell, D.T.; Schwartz, R.D. and Sechrest, L. 1966. ***Unobtrusive Measures.*** Chicago, Illinois: Rand McMally.