TOP

From: Bulletin of International Test Commission, 1989, No. 29, 67-9

QUESTIONABLE ASSUMPTIONS IN TEST CONSTRUCTION

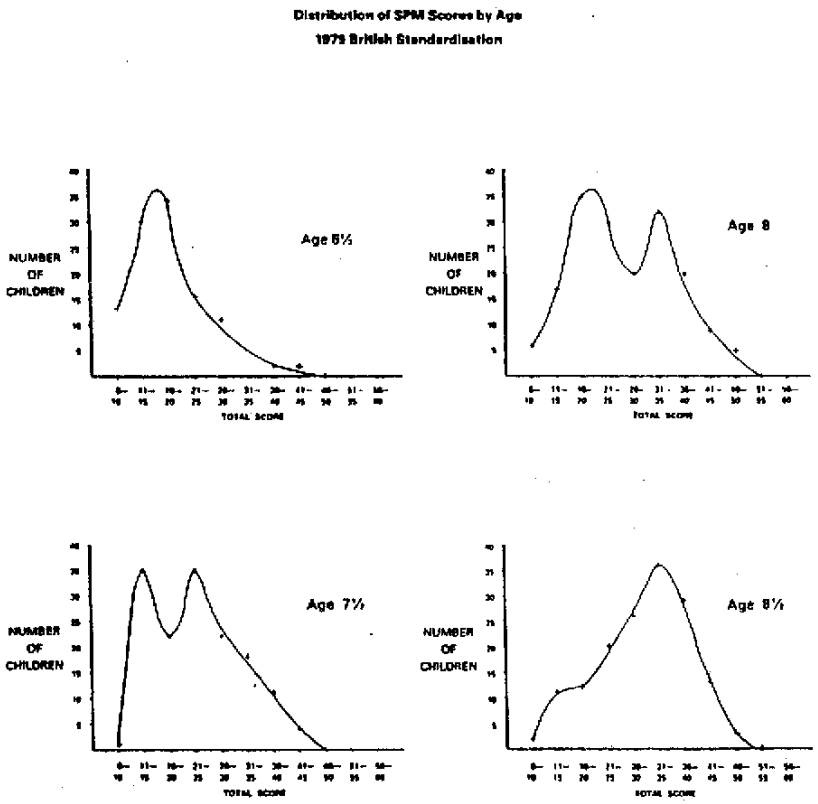
John Raven
Consultant on Educational and Psychological Assessment,
30 Great King Street, Edinburgh, Scotland

The following assumptions, commonly made by psychometricians and test users, are challenged.

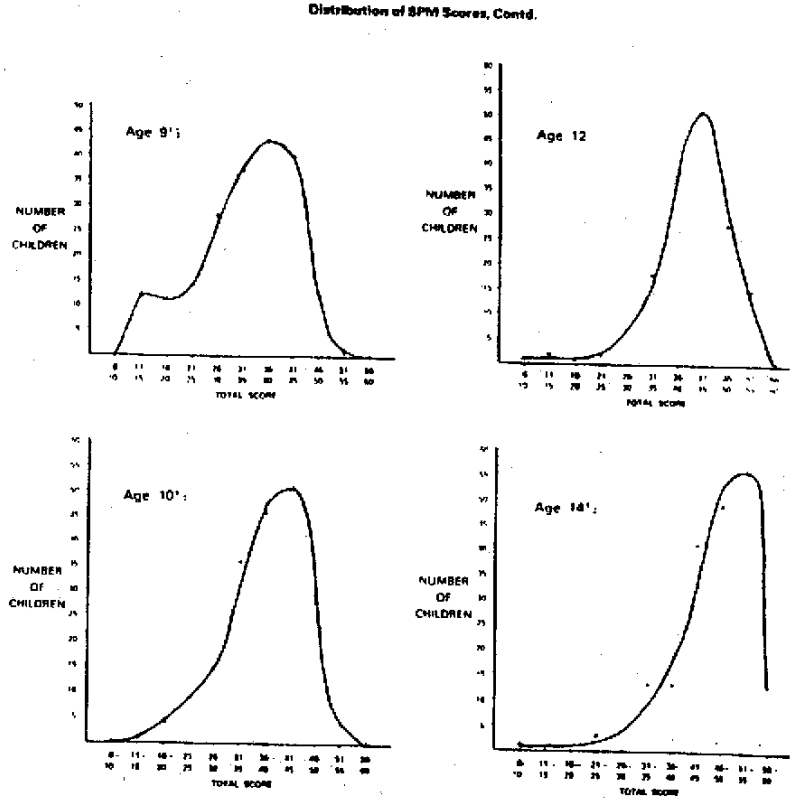
- Test scores should be "normally" distributed.
- Tests should be value free.
- Tests should have high internal consistency.
- Psychometricians should not be concerned with respondents' social or political beliefs.
- People can be described independently of the situation in which they find themselves.
- We are most likely to find a parsimonious way of handling individual differences by seeking a small number of variables which will summarise the bulk of the variance in human concerns and abilities.
- It is important to find ways of making fine discriminations between people along a small number of dimensions.
- The practice of making descriptive statements about people is unscientific.
- Making fine discriminations on a small number of dimensions is "objective", whereas a statements about people's distinctive characteristics, or about the effects of educational and social programmes and policies, is not.
- To be useful for scientific or practical purposes, any trait which is indexed should be stable.

1. THE ASSUMPTION THAT SCORES SHOULD BE "NORMALLY" DISTRIBUTED

Three points need to be made about this assumption. The first is that, in actual fact, test scores are not, typically, distributed according to a "chance" Gaussian curve. The curves obtained in the 1979 British standardisation of the Raven Progressive Matrices are shown in graphs 1 to 9 and enthusiastic readers can trace the curves from the earlier Irish[1] and British standardisations from the



Graphs 1 - 4



Graphs 5 - 8

references in the test *Manual*[2].

According to Thorndike (in a personal communication), the distributions for the sub-tests of the Stanford Binet are likewise almost always bi-modal. However, this is concealed by first reading off percentile scores corresponding to each raw score and then converting the percentiles to deviation-IQs. Given that the procedures for combining scores and calculating eligibility for Special Education assume "normal" distributions, this would seem to be a highly dubious practice - particularly as it reinforces the use of Mental Age and IQ. These concepts carry many unacceptable overtones. For example, the concept of Mental Age (with its embedded assumption that "children of the same mental age are more like others of the same mental age than like children of their chronological age") reinforces the practice of accelerating and retaining children in school instead of stimulating a demand to diversify educational programmes in such a way as to cater for a wide range of children who have very different concerns, talents, abilities, stores of information, and patterns of previous experience. IQs, for their part, are widely assumed to be stable over time and to have pervasive predictive validity.

One can, of course, tinker with the number of items at each level of difficulty to make the curves come "right" for any one group, but, almost by definition, if they are "right" for one group they cannot be right for all sub-groups - including all age and ethnic groups having different mean scores - and all total groups - which will, by definition, have different proportions of each age and ethnic group. The overall Matrices scores for primary school pupils is, for example, as shown in Graph 10.

However, whatever the theoretically best distribution of natural or "uneducated" abilities, the implication of Ben Bloom's writing on mastery learning is that a "normal" distribution of *educated* abilities is an indictment of the educational system. If it is important for children to learn something, then it is iniquitous - and evidence of a very serious failure on the part of their *teachers* - if they have not learned it. If it is not important for them to have learned it, it is disgraceful that they should have had to waste their time being subjected to the indignity and despair of having had to sit through classes lost in a fog about what is going on and often being physically punished for non attainment ... and again, I would have thought, legally actionable evidence of *teacher* failure.

What this means is that the only justifiable distribution of educated abilities is as in Graph 11, where pupils with different needs have been brought to different levels of mastery. (The real

problem is that the tension discussed in this paragraph tells us that schools and testing do not exist mainly to provide and enhance education but to discriminate between people in a way which legitimises social inequality in position and status.)

Another good reason for *not* trying to develop tests which have "normal" distributions is that they are rarely suited to the task of demonstrating that the educational system as a whole, or any sub-section of it, is *not* achieving important goals. One cannot get a 'normal' distribution across the entire population when only a few teachers are fostering the relevant talents. (This is one reason why it has so rarely been demonstrated that schools, in general, stunt the development of qualities like initiative and the ability to work with others.)

A corollary of this is that few of the tests which are currently available are suited to the task of demonstrating that the educational system - or any given teacher - is (or is not) producing a wide variety of different types of excellent student. To do so, one would need to lay current concepts of discrimination on their side and use tests which identify *what* each student is good at instead of whether he or she is good at performing a very limited range of tasks selected by the teacher or evaluator.

2. THE ASSUMPTION THAT TESTS SHOULD BE VALUE-FREE

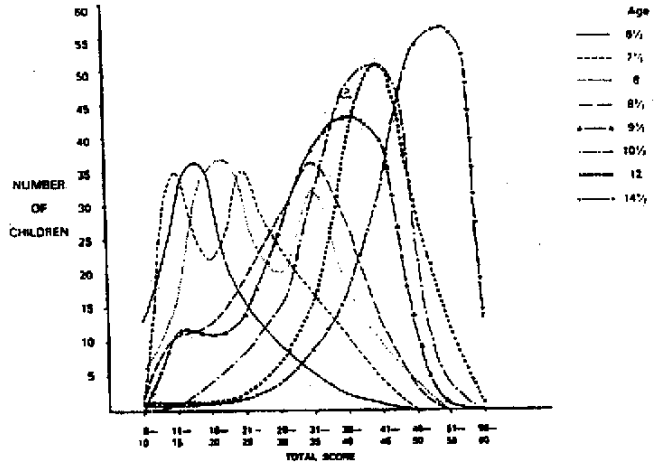
It is easiest to challenge the notion that psychological measures should be value-free by analysing the nature of some of the characteristics which it is generally agreed that it is vital for educators to foster. These include leadership, the ability to understand and influence society, and initiative. For illustrative purposes, I will focus on the last of these.

To take a successful initiative, people have to be self-motivated. Self- starting people must be persistent and devote a great deal of time, thought and effort to an activity. They need to initiate innovative action, monitor the effects of that action, and learn from those effects more about the problem they are trying to tackle, the social, political and environmental context in which it is situated, and what is effective and ineffective about the strategies they are using. To succeed, they must anticipate obstacles in the future and invent ways of circumventing or overcoming them. They need to build up their own, unique, set of specialist knowledge. They will have to get help from others. More often than not, it will be necessary to establish coalitions with others to gain control over social and political forces which would otherwise deflect them from their goals.

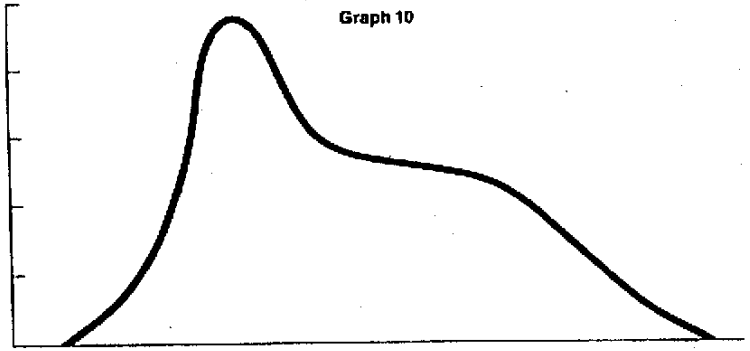
Graphs 9

— 1979 BRITISH STANDARDISATION OF SPM —

SUPERIMPOSED AGE CURVES

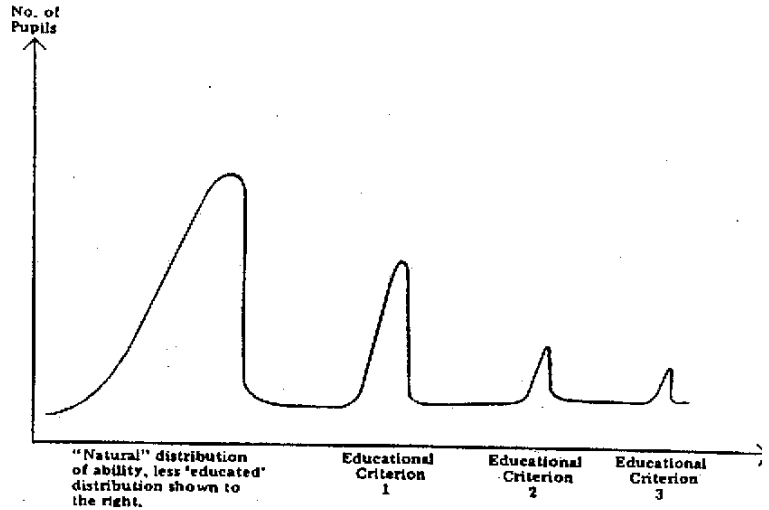


Graph 10



Raw Scores for 3466 Children on S.P.M

Graphs 9 - 10



Graph 11

The same is also true of other important qualities like the ability to identify and solve problems, leadership, and the ability to work with others.

It follows from what has been said that one must know an individual's values, preoccupations, or intentions before one attempts to assess his or her abilities. Important abilities demand time, energy and effort. As a result, people will only display them when they are undertaking activities which are important to them. This model can be made more concrete by reference to Grid 1 on which a selection of potentially valued activities have been listed across the top and competencies which may be brought to bear to undertake them effectively down the side.

The above analysis also implies that it does not make sense to attempt to assess separately the cognitive, affective, and conative[3] components of an activity. One cannot meaningfully assess 'the ability to develop better ways of thinking about things' independently of the pleasure which the person concerned derives from doing so, and his or her determination to make glimmering insights explicit. *These affective and conative components are an integral part of what we mean by "cognitive ability"*[4]. Not only do the three components interpenetrate: if the behaviour in question - the initiative - is to be successful, these components must be in balance. Both determination exercised in the absence of understanding, and the converse, are unlikely to make for success.

These observations are in sharp conflict with many traditional canons of psychometry. I have argued that one cannot assess abilities independently of values. This means that it is essential to adopt a two-stage approach when assessing competence. We must first find out which types of behaviour someone values, and then, and *only* then, assess his or her ability to bring to bear a wide variety of potentially important cognitive, affective and conative behaviours to undertake the activity effectively.

It is important to emphasise that the widely held view that one can use one set of scales to assess values and another, independently, to assess knowledge, skills, abilities or competencies, simply does not make sense. The latter will only be developed and displayed in relation to the former.

3. THE ASSUMPTION THAT TESTS SHOULD HAVE HIGH INTERNAL CONSISTENCY

Our example - initiative - also highlights another way in which the assumptions on which the dominant measurement paradigms in psychology and education are based fail to engage with important

aspects of competence. Conventional psychometric theory place great stress on internal consistency or factorial purity. Score derived from tests composed of items which do not correlate wit each other are said to be meaningless. Yet it would seem from ou example that this assertion is not necessarily correct. People initiatives are more likely to be successful the *more* independer and different things they do in the course of undertaking activitie they care about. For example, they are more likely to be successfi if they re-conceptualise the problem, obtain the help of other persist over a long period of time, and so on. Yet their inclinatio and ability to do any one of these things in pursuit of their goals i unlikely to be closely related to their inclination and ability to d other things. Furthermore, if they do any one of them particularl well, it will, to some extent, compensate for their failure to d others.

It follows from the observations made in the last paragraph tha if we are to assess such qualities as initiative, instead of trying t develop measurement tools which are as internally consistent a possible, we need to try to develop *indices* made up of items whic are as little correlated with each other as possible[5]. (This i actually not so heretical as at first sight it appears, because it i standard practice to make use of multiple regression equation which involve summing over maximally independent variables t obtain the best prediction.)

4. THE ASSUMPTION THAT THE PSYCHOMETRICIAN IS NOT CONCERNED WITH THE RESPONDENT'S SOCIAL AND POLITICAL BELIEFS

Another disturbing insight which follows from our analysis c "initiative" is that value-based cognitions of social processes ar central to competent behaviour and need to be documented in an meaningful assessments of competence.

Behaviour is very much determined by such things as people's beliefs about how things *should* be done, who should relate t whom, and about what. It is influenced by people's perceptions o roles - by what they think it is appropriate for someone in thei position to do, by what they think other people expect them to do and by how they think other people will react to their behaviour. I is determined by their understanding of what is meant by terms lik 'management', 'participation', 'majority decision-taking', 'managerial responsibility', 'wealth', and 'democracy'. And it is determined by their beliefs about how society should be managed.

The disturbing conclusion is that, if we are to assess competence

Grid 1

A Model of Competence

Valued styles of behaviour

	Achievement				Affiliation				Power	
	Doing things which have not been done before.	Inventing things.	Doing things more efficiently than they have been done before.	Finding better ways of thinking about things.	Providing support and facilitation for someone concerned with achievement.	Ensuring that a group works together without conflict.	Establishing warm, convivial relationships with others.	Establishing effective group discussion procedures.	Ensuring that group members share their knowledge so that good decisions can be taken.	Articulating group goals and releasing the energies of others in pursuit of them.	Ensuring effective compliance with one's demands.

Cognitive

Thinking (by opening one's mind to experience, dreaming and using other sub-conscious process) about what is to be achieved and how it is to be achieved.

Anticipating obstacles to achievement and taking steps to avoid them.

Analysing the effects of one's actions to discover what they have to tell one about the nature of the situation one is dealing with.

Making one's value conflicts explicit and trying to resolve them.

Consequence anticipated:
Personal: eg "I know there will be difficulties, but I know from my previous experience that I can find ways round them."
Personal normative beliefs: eg "I would have to be more devious and manipulative than I would like to be to do that."
Social normative beliefs: eg "My friends would approve if I did that"; "It would not be appropriate for someone in my position to do that."

Affective

Turning one's emotions into the task; admitting and harnessing feelings of delight and frustration; using the unpleasantness of tasks one needs to complete as an incentive to get on with them rather than as an excuse to avoid them.

Anticipating the delights of success and the misery of failure.

Conative

Putting in extra effort to reduce the likelihood of failure.

Summoning up energy, determination and will-power.

Persisting over a long period, alternately striving and relaxing.

Habits and experience

Confidence, based on experience, that one can adventure into the unknown and overcome difficulties. (This involves knowledge that one will be able to do it plus a stockpile of relevant habits).

A range of appropriate routinised, but flexibly contingent behaviours, each triggered by cues which one may not be able to articulate and which may be imperceptible to others.

Experience of the satisfactions which have come from having accomplished similar tasks in the past.

in any meaningful way, it will be necessary to assess these, essentially political, beliefs.

5. THE ASSUMPTION THAT PEOPLE CAN BE DESCRIBED INDEPENDENTLY OF THE CONTEXT IN WHICH THEY FIND THEMSELVES

Although the way in which people define the situation in which they find themselves has a marked effect on their behaviour, that context has other direct and indirect effects. It influences their behaviour directly through the constraints which it places on what they can do, and it influences it indirectly through the concepts, understandings and competencies which they are able to practise and develop.

It therefore emerges that, if one wishes to assess competence, it is necessary to assess both the perceived and the actual institutional context in which it occurs. It is either meaningless or wildly prejudicial to say that people lack the ability to do something which they have never had the opportunity to practise doing because they may be able to do it extremely well once they have had a little practise. It is also prejudicial if the situation fails to elicit the behaviour or if the task fails to engage the motives of the person being assessed. The only way out of the dilemma is to make assessment of the context and previous opportunity to learn part of the assessment of the individual.

6. THE ASSUMPTION THAT IT IS REALISTIC TO THINK THAT THE WAY FORWARD IS TO BE FOUND BY STRIVING TO FIND A SMALL NUMBER OF DIMENSIONS WHICH WILL SUMMARISE THE VARIANCE IN HUMAN CONCERNS AND ABILITIES

It is easiest to illustrate the problems which current assessment paradigms pose for educational evaluation and assessment from the work of one school class whose work was described in detail in our book *Opening the Primary Classroom*[6]. In this class, most of the students' education took place in the course of interdisciplinary projects. These projects were very thorough-going. The students, as a group, carried out original investigations in the environment. Their work inside their classrooms formed an integral part of these investigations. Within these overall projects, many students had personal projects, distinctive areas of specialisation, and distinctive roles.

What was most striking about the approach was the teacher's distinctive concerns. She was not preoccupied, as were most teachers, with course work; with covering a syllabus. Her attention

focused on the competencies which she hoped to help her students to develop through the activities they carried out. These competencies included reading, writing, spelling and counting. But they also included communicating, observing, finding the information which was needed to achieve goals (such information often having to be collected by observation or by talking to people rather than by reading books), inventing, persuading, and leading.

It is easiest to begin our discussion of the problems which this educational process poses for assessment by reviewing those aspects which are closest to the more widely discussed and assessed goals of education.

In the course of his environmentally-based project work, one student had become an expert on the distribution of different species of butterfly in the locality, their life cycles, and their relationship to their habitats. Another had become an expert on the history of a particular agricultural implement: he had related changes in the implement to a continuous - and apparently autonomous - series of improvements in the design itself (themselves dependent on improvements in the processes used to manufacture steel) on the one hand, and to changing patterns of agriculture on the other. Another student had become an expert on the relationship between improvements in that implement, the pattern of land use it demanded and facilitated, and changes in the social structure of the community. Yet another had investigated the current social structure of the area - who knew whom and what they talked about. Others had studied changes in the architecture and layout of the village and the occupations of its inhabitants.

The problems which these accomplishments pose for conventional measurement paradigms are almost insurmountable. The students' specialist knowledge simply would not show up on traditional attainment tests - indeed these students would get *low* scores on such tests because they would not have mastered the required knowledge. It may be thought that this comment only applies to "exceptional" information but, as this teacher noted, even her English teaching was seriously threatened by tests which failed to pick up either her students' ability to use structure and what was said between the lines to obtain information which was of value *to them* or their ability to use innuendo and juxtaposition of ideas to convey important images and ideas. To do justice to the students, it would be necessary to administer a series of individualised tests which would tap each student's specialty.

However, these are the least of the problems which this work poses for measurement. More important than the unique store of

specialist knowledge built up by the first student mentioned above was the fact that he had developed a selection of the *competencies* required to be a scientist. Among other things, he had learned to be sensitive to the cues which tell one that one has an unresolved problem; he had developed the tendency to try to make glimmerings of insight on the fringe of consciousness explicit; he had learned to *invent* ways of making observations; he had learned to notice things which no one had noticed before; he had learned not only how to find information in journals, but also how to use what he did find to stimulate that kind of lateral thinking which is required to make use of the information that is obtained; he had learned to solicit and make use of the ideas of his fellow students and 'ignorant' people in the community; he had learned to write to, telephone, and visit university lecturers who were interested in the same problem and he had spoken to them as equals; he had sharpened up his ideas by sparring with them; he had learned that he had a right to ask new questions and not merely answer other people's; he had learned that he could both ask and answer questions; he had learned to tolerate the frustrations which are inherent in trying to find better ways of thinking about things; and he had learned to invent ways of summarising his data and communicating it to others - and not just in writing.

The competencies listed in the last paragraph are a sub-set of the competencies which cumulate to result in the effective pursuit of valued goals, and which can, to a degree, be substituted one for another[7]. We have come upon them here in connection with discipline-based studies, but we could equally well have encountered them as a result of examining other activities which people might value and be motivated to undertake effectively. But, pursuing the academic- discipline-oriented line of enquiry on which we have embarked, it is important now to note that the second student mentioned above had developed a different sub-set of these self-motivated preoccupations, sensitivities, thoughtways and perceptions in the course of undertaking an original historical study. The third had developed a similar - but by no means identical - selection of the competencies needed to be an excellent sociologist of one kind or another. And so on for the other students.

If our traditional assessment procedures are unable to cope with the problem of idiosyncratic, specialist, high-level, new knowledge, they are even less able to document the growth of the subtle skills, motivated habits, thoughtways and preoccupations which go to make up the repertoire of the competent scientist, historian, sociologist, photographer, reporter, cook, or mother.

But even this does not exhaust the problems which th educational process in which these students were engaged pose fo assessment. The students had worked as a group. They ha developed specialised roles in that group. In the process they ha developed the competencies needed to function effectively in thos roles. One student had become good at co-ordinating the activitie of others. Another at putting others at ease and smoothing ove interpersonal difficulties. Another a negotiator. Another a presenting the results of other people's work to external visitors - communicator rather than an original researcher. And so on. I the course of undertaking these activities, all students learned t communicate, to invent, to make their own observations, to worl with others, to lead and to follow.

These competencies defy conventional measurement. This is o the greatest importance. Without means of assessing thes qualities, even students who have come through such educationa programs cannot know that they are different from students wh have come through other educational programs. Still less can the identify the ways in which they are different from them: the cannot know that they think differently, see things differently, hav different priorities, tend to work differently with others or that the can do different things. Without means of assessing these qualities, teachers cannot build on the competencies which have been fostered in one project in the course of the next one. Students cannot ge credit for the talents they have developed when they try to get a jo or enter a course of further education. Teachers cannot get credi for having fostered these competencies in accountability exercises. As a society, we cannot prevent people who do not possess important high-level concerns and qualities like those mentioned being appointed to influential positions. But, perhaps worst of all, the absence of means of measuring these qualities limits our conception of what education *is*. In the current scheme of things, even the word 'academic' fails to denote activities in which people observe, think, find better ways of thinking about things, or communicate effectively (as distinct from following rules of grammar which, as the Bullock committee[8] has noted, have never been shown to relate to effective communication).

Before moving on, it is also worth noting that the fact that we have been able to make these observations shows that the measurement problem must, in principle, be solvable. The key features of what we have done are that we have: (i) observed students as they were undertaking tasks they cared about, (ii) taken the trouble to record the multiple and substitutable competencies they displayed whilst undertaking those tasks, (iii) made use of a

descriptive framework rather than one grounded in 'variables' to identify the effects, and (iv) distinguished between students in terms of the activities they cared about and the competencies they displayed whilst undertaking those tasks, rather than in terms of their scores on a small number of 'dimensions'. The central message of my publications on assessment (e.g. *Competence in Modern Society*[9], and my paper in Black and Dockrell, 1988[10]) is that it is possible to build an alternative measurement paradigm based precisely on what we have done here.

In concluding this section, it is important to emphasise that what these observations show is that there is no way in which it would be possible to give a reasonable account of this teachers' effectiveness or her effect on her pupils, or a reasonable account of the competencies possessed by any individual student, if one limited oneself to a small number of variables - or, indeed, if we employed any variable-based model.

7. THE ASSUMPTION THAT IT IS IMPORTANT TO FIND WAYS OF MAKING FINE DISCRIMINATIONS BETWEEN PEOPLE ALONG A SMALL NUMBER OF DIMENSIONS, RATHER THAN CRUDE DISCRIMINATIONS ACROSS A LARGE NUMBER OF CATEGORIES

Psychologists have aided and abetted teachers and managers who believe (or who find themselves in positions which lead them to act as if they believed) that there is some real value in making fine discriminations between people's ability along a few dimensions.

In fact, few of the tests which are used are able to support such discriminations: they do not have the discriminative power which the practice of transforming raw scores to extended deviation scores has led users to believe that they can make, they are so unreliable that the discriminations cannot have any validity, and they lack sufficient construct and predictive validity.

Before supporting these assertions, it may be useful to illustrate some of the uses which I believe it is important to challenge. In the course of discussions I have had in many parts of the United States, I have repeatedly come across scenarios like the following: the state legislature had laid down that there would be "gifted" programmes in the schools and had acknowledged that there were 6 different types of giftedness. They had decreed that giftedness in any one of these respects would make a child eligible for entry to the programmes. The psychologists concerned were then faced with the problem of finding ways to select about 5% of the population on the basis of the six alternative criteria. These were: academic

success, a high Matrices score, excellence at Music, Art, or Sport, a high Parental rating for Leadership. What were they to do? Wh they did was try to identify the top 5%th of a percentile on ea criterion! I found myself being asked to specify precisely what t appropriate Matrices cut-off score would be. (Absurd though th part of the process was, its absurdity competes only with wh happened next - for, having admitted all these different types gifted student, those concerned made no attempt to ensure that t educational programme which were offered catered for diversi Instead, they offered all pupils an accelerated academic programm

The use of attainment tests to determine pupils' futu education and entire careers is widely accepted as legitimate. Mo users believe that there is a big difference between the ability of top grade student and one who fails. Yet Spencer[10] found that t raw score difference between the two on typical academic tests typically no more than 8 marks. This small difference is th transformed into a difference between "above 77%" and "belo 50%".

As far as the generalisability of the statements which are ma on the basis of attainment tests is concerned, it is of more tha passing interest to note that Wolf[11] showed (1) that whether or n a student has 'mastered' particular operations depends very muc on the way the questions are asked and the tasks are set (thus, fi example, the proportion of students who are 'able to undertak simple divisions' is dramatically different if the questions are pose in the form of 112/7 rather than in the form 112÷7 = ? and th dependence of the answer to the question of whether the 'criterio has been reached on the way in which the task is presente becomes much greater when the criterion consists of being able t undertake any kind of task remotely resembling real-life), (2 examiners commonly do not agree even on whether a particula performance should be assigned to the top or bottom *halves* of th distribution, let alone in their more detailed ratings, (3) the sam examiner will usually allot very different marks to the sam performance on different occasions, and (4) examinees perform th same task in very different ways on different occasions. He conclusion is that the *constructs* to be indexed have very littl generalisability and that re-test reliabilities are extremely low even when the measures have high internal consistency. (Indices o internal consistency are, of course, typically presented as acceptabl surrogates for re-test reliabilities.)

Actually, it is not really necessary to cite such technicalities to call the construct validity of attainment tests into question. To take

an example, there is no sense in which temporary knowledge of a smattering of out-of-date scientific facts can be said to be a valid index of scientific knowledge - for such a label gives the impression that the knowledge assessed is in some sense a representative sample of the whole domain of scientific knowledge. Still less can a score on such a test be described as an index of the ability to think scientifically, the ability to keep up to date in a specialist field, the ability to find information relevant to problems one encounters, or the ability to make observations which aid in solving those problems.

If the construct validity of academic assessments has rarely been carefully investigated, the same cannot be said of their predictive validity. Unfortunately, Ingenkamp's[12] review of the extensive literature reveals that, not only are the results of these investigations none too positive, the selection of appropriate criteria against which to validate achievement tests is problematical, to say the least.

Although employers regularly use attainment tests to select employees, the Hunters[13] meta-analysis of some 90 studies shows that their predictive validity even to success in *training* is only about ·2. This is significantly, although not dramatically, less than the predictive validity of intelligence tests. This is, however, not the end of the story, because Berg's[14] work shows that the correlation between earlier and later 'educational achievements' - even in the same subject - drops to zero if the subsequent courses are taken some time after leaving school, and especially if those courses are really required to improve job performance.

Hunter and Hunter[15] minimise the inability of educational attainment tests to predict more than 1% of the variance in occupational *performance* (over 425 studies) by saying that the criterion measures which are available are unreliable and that, in any case, too many other factors come into operation.

On reflection, however, how *could* traditional tests be expected to predict occupational success with any validity? In the first place, it is almost inconceivable that any small number of tests should have high predictive validity: what a person will do depends on a whole constellation of factors which come into play in the particular circumstances in which he or she finds him or her self This suggests that we should be more concerned to cover the whole domain of an individual's talents and abilities and to describe the situation in which he or she finds himself than to get an accurate score on a few variables. Secondly, any one occupational group contains within it a wide variety of people who do very different

things. Some psychologists manage factories, others review research, others do original research, others treat patients, and others write computer programmes. As Berg[16] has emphasised occupational groups are to be understood, not as psychological phenomena, but as sociological phenomena, the function of which is to regulate competition. By maintaining 'qualified' entrants in short supply, their incumbents are able to achieve scarcity value Berg's data show that entry qualifications are raised when the number of 'qualified' entrants is about to increase, and Folger and Nam and O'Toole[17] have shown that this occurs despite there being no change in the tasks undertaken, or the skills exercised, by the group concerned.

What this means is that the attempt to make fine discriminations between people for practical purposes cannot be justified. In the next section, I will argue that, instead of pursuing this will-o-the-wisp, we would be better to pursue the goal of getting a rough fix on people's overall pattern of motives, beliefs and competencies and the environments in which they are - and are likely to be - placed.

8. THE ASSUMPTION THAT THE PRACTICE OF MAKING DESCRIPTIVE STATEMENTS ABOUT PEOPLE IS UNSCIENTIFIC

As an undergraduate, I was markedly influenced by Eysenck's *Uses and Abuses of Psychology*[18]. This argued that category-based verbal descriptions of people are unscientific and that what psychologists needed to do was to first make a dimensional structure explicit and then arrange people along (give them scores on) each of these dimensions.

It is of more than chance significance that the first computer programmes used to perform factor analyses in psychology were programmes which physicists had developed to resolve fields of forces: the *variable*-based model which lay behind the factorial/dimensional model of human behaviour is indeed drawn from physics and based on physicists' models.

The kind of model that physicists tend to work with is known to any schoolboy who can tell one that the behaviour of a projectile is best described by some such equation as:

$$s = ut + \tfrac{1}{2}ft^2$$

(the distance travelled at a particular time is determined by the initial velocity multiplied by the time elapsed plus half the acceleration multiplied by the square of the elapsed time).

The "dimensionally" oriented psychologist's model is analogous. Such a psychologist may, for example, assert that the amount of leadership which someone will display will be a function of that person's scores on such variables as extroversion and intelligence.

The physicist's model is not, however, the only one which is available, even to schoolboys. Many know that chemists have found a very different type of equation to be the most useful. Chemists argue that substances and the environments in which they are placed can best be described by listing the elements of which they are composed, and the relationship between these elements. The descriptors (elements) are drawn from a large set known to all chemists. The elements which are not present do not need to be listed. The behaviour of a substance in a particular environment is then described by equations which make it possible to describe transformations as well as monotonic combination:

$$Cu + 2H_2SO_4 = CuSO_4 + 2H_20 + SO_2$$

(Copper plus sulphuric acid yields copper sulphate, water and sulphur dioxide).

The first and most important point I want to make here is that it follows from these observations that it would *not*, in principle, be "unscientific" for psychologists to adopt a chemist-type, descriptive, paradigm. Human beings might best be described by listing their values and the components of competence which they show a spontaneous tendency to display when they encounter particular environments. At the simplest level, this could be done simply by placing check-marks in the appropriate cells of Grid 1. The environments in which people were observed could be described in an analagous way. Such statements might take the following form (the symbols which are used are exemplary only, and should in no way be taken to suggest that we have developed even a preliminary version of a more complete table of 'human elements'):

$$Ach_4Pow_3;\ Auth_4PartCit_2;\ NuP_4HostP_3;\ DP(T)_1$$

Such a statement might be interpreted to mean that the individual showed a spontaneous tendency to display four components of competence in pursuit of achievement goals, and three in pursuit of power goals. Four items contributing to the set dealing with authoritarian perceptions of society, and only two of the set dealing with participatory citizenship, were endorsed. Four aspects of the environment were supportive of the individual's goals: the manager modelled achievement behaviour but did not delegate, encourage participations, nor create developmental tasks for his

subordinates. There was 'hostile press' from other people in th individual's environment. Concern with efficiency and effectiv leadership were scorned. The task which the individual was set ha little developmental potential: it was a routine task which prevente the person concerned from developing perceptions and expectation appropriate to innovation.

If the equation were written in some way which permitted (movement, one would conclude that the individual would be likel to become frustrated and lose motivation to engage in achievemer and leadership behaviours.

In fact, of course, such summary statements could be filled ou usefully in a great deal more detail. Once could identify exactl what type of achievement or power behaviour the individual thoug it was important to engage in; one could identify exactly whic competences were brought to bear in pursuit of each; one coul identify the particular perceptions and expectations whic encouraged and prevented the person concerned from engaging i such behaviour; one could say more about the role models to whor he or she was exposed by managers, colleagues and subordinates and one could say more about the tasks set and their probabl effects on the person's future development and motivation.

Before leaving this discussion, it is worth reiterating that suc *statements* about people *and the environments in which they ar placed* enable us to handle the transformational processes whic occur in homes, schools and workplaces.

One final observation may be made about the methodolog advocated here. The crucial component in what we have been doin is that we have been mapping and sampling relevant *domains* o beliefs and expectations. This is no routine handle-cranking concurrent-validity-drive, activity. Quite the opposite: it can onl be carried out effectively in the light of a thorough understanding o the area one is dealing with. The need is, therefore, not so muc for a new *methodology* as for a clear expectation that scientist should devote time to what is, after all, the crucial phase of an scientific enquiry worth the name.

9. THE ASSUMPTION THAT MAKING FINE DISCRIMINATIONS ON A SMALL NUMBER OF DIMENSIONS IS "OBJECTIVE" WHEREAS A CLUSTER OF CRUDE STATEMENTS MAPPING THE ENTIRE FIELD IS NOT

The most commonly encountered objection to any change in assessment procedures is that the alternatives are 'subjective'. However, current assessment procedures cannot be described as

"objective" in anything but a trivial sense. How objective is a picture of a student or employee which ignores the most important concerns and priorities of that person, the talents which he or she is able to exercise in pursuit of them, and the situational constraints which prevent him or her from deploying their talents in the situation in which the observations have been made? Yet such one-sided descriptions are precisely what we are most commonly offered by schools and assessment agencies - and justified by reference to the supposed objectivity of the information which is provided.

In fact, value-based decisions are built into *all* evaluations. In this case, they have been obscured by a facade of pseudo-science. What is assessed and reported is *primarily* dependent on the preoccupations, values and purposes of those who construct and select the tests that are used, and not on the qualities of the person being assessed. Value-free assessment, value-free evaluation, and, indeed, value-free education, are self-contradictory concepts; oxymorons. Not only is value-free assessment an impossibility: the assertion that assessment should be value-free is itself a value-based statement asserting a value-based criterion for the evaluation of assessment procedures which has the effect of concealing, even if unintentionally, the values of the person making it. The sooner this fact is grasped *and built into our concepts of assessment, evaluation, and education*, the sooner we will be able to make progress.

A more important limitation of current concepts of objectivity is, however, the fact that the qualities which someone will display are vitally dependent on the situation in which he or she is placed, and on the opportunities which he or she has had to develop the qualities which are being assessed. It is difficult to demonstrate acceptable forms of initiative in a Latin class. The selection of the Latin class as an appropriate context in which to rate 'initiative' amounts to a statement about the *kind* of initiative which is valued by the assessor. Consequently, not only are assessments of initiative made in Latin classes unlikely to have much predictive validity to other situations: they are unlikely to reflect the capacity of those being rated to display initiative in other situations. This does not invalidate the concept of initiative; it simply shows that its display and recognition is dependent on the values of both the person being assessed and the assessor. Thus, insistence on standardised test items and situations, while perhaps conducive to replicability in behaviour, are not after all conducive to the objective assessment of competence. Even if the candidate values the task set or the benefits which might be obtained by performing it 'well', the situation may preclude the levels of persistence, persuasion, and

planning which are crucial to any successful initiative. Likewise differences between candidates may tell us more about th opportunities they have had in the past to practise taking initiativ in those kinds of situations, than they tell us about what thos same candidates are objectively capable of doing in those situation - or habitually do in other situations which are of concern to then Thus, differences between candidates in typical assessment cente tasks - such as persuading unwilling colleagues to build a bridg across a river to convey tanks to murder innocent "enemies" usually tell us more about whether they have had relevant value and experiences than they do about what the candidates would do they had had relevant experience, or what they do when trying t undertake activities which they care about and of which they hav experience. They may have neither a predilection nor the capacit to persuade others to build a bridge which someone else has tol them to get built to perform a task which they abhor, but they ma be excellent at inventing new materials for building bridges an persuading others to provide relevant information. A demonstratio of failure to record this other fact would seriously challenge th assessment center's claim to objectivity.

The observations made in the last paragraph mean that failur to document (1) the context in which observations are made and th behaviours which that situation tends to elicit, fails to elicit, an suppresses, (2) the opportunities which the person being assesse has had in the past to develop relevant or alternative competencies (3) what the candidate can do if he or she is able to work for ar extended period of time at a task which he or she cares about, an (4) what kinds of situation would engage the values, motives an talents of the candidate, is therefore a serious indictment of th objectivity of any assessment process. Yet few of the curren procedures which claim to be objective provide any of thi information. The typical disclaimer - that the test score reports onl what the person being assessed did on a specific test at a particula time - shifts the responsibility for any injustices which may resul from the assessment from the (professional) assessor to the (lay) user. But it does not make the process any more objective. And i is almost certainly an infringement of professional ethics.

Similar comments can be made with even more force about evaluators' claims to evaluate projects and policies with "objectivity" because they have followed the recommendations of the Joint Committee[19] and used only "reliable and valid tests". In our own work, we have shown that, if one asks "What effect is this teacher having on her students?" (instead of "What effect is this teacher having on her pupils' performance on these selected tests which

cover only a very limited range of outcomes which do not include the idiosyncratic information communicated by this teacher) one finds that different teachers have dramatically different effects on their pupils' values, concerns, priorities and patterns of competence. While some teachers have outstandingly positive effects (which do not show up on conventional tests), most teachers have very negative effects on most of their pupils. In this context, how much credibility can be placed on the claim made by traditional evaluators (who record neither the outstanding benefits conferred on their pupils by some teachers nor the harm done by others) that their results are "objective" because they could in principle be replicated by other researchers?

10. THE ASSUMPTION THAT THE TRAIT BEING INDEXED IS STABLE (UNINFLUENCED BY THE ENVIRONMENT)

There are two reasons why it is very important to surface and re-examine this assumption. On one hand, many teachers and others tend to assume that if a quality can be measured it is stable. Conversely, others (like Jim Flynn[20]) assume that if measures are not stable, then little of real meaning is being assessed.

Neither of these positions is correct. I have already shown that many characteristics will only reveal themselves and develop if the person concerned is able to undertake activities which he or she cares about in an appropriate environment. The environment releases, enables people to express and perfect, and transforms, inherited characteristics.

The same is true of the characteristics of wheat. Not only do the heights and yields of different strains of wheat change when they are grown in different environments, the *rank order* of both height and yield changes - and in such a way that the correlation between height and yield - between each and any third variable - changes dramatically.

What these comments imply is that we need to substitute a concern with construct validity for our concern with even re-test reliability when we are developing assessment procedures.

IMPLICATIONS

In this paper, I have challenged ten common assumptions in test construction. Elsewhere[21], I have discussed the consequences of failing to develop measures which avoid these assumptions in some detail. It is sufficient simply to list them here.

* Because it will remain impossible to develop the tools which

teachers require to implement the individualise competency-oriented, educational programmes which fost multiple and alternative talents which are required to achie the main goals of education, and because teachers and pupi will continue to be unable to get credit for the accomplishments, we will continue to squander some two-thirc of the money we devote to the educational system[22].

* Because it will remain impossible to identify most people talents, they will remain underdeveloped and underutilise This will result in considerable cost both to themselves an society. For the same reason, promotion into influenti positions in society will remain essentially random so far as th talents needed in those positions is concerned.

* Because it will remain impossible to mount broadly base evaluations of educational policies and programmes, thos evaluations - and all policy discussions based upon them - wi continue to be misleading and the consequences for societ socially dysfunctional[23]. For the same reason, "Back- to-basic attainment testing" evaluation and accounting to guid education (as in the US and UK) will deflect schools from thei goals, prevent most people developing their talents, and promot the wrong people into influential positions in society. Without a change in our framework, psychologists will be unable to offe society any alternative system of accountability. The choice wil remain: "3rs or nothing".

* Because it will not be possible to give public servants credit for possessing and deploying high-level competencies, effective staf appraisal and accountability exercises will not be implemented. This will mean that it will continue to be impossible to ensure that public servants seek out relevant information and take appropriate innovative action in the public interest[24].

* Because it will not be possible to identify the real talents possessed by differing ethnic and socio-economic groups (and thus offer equity in diversity in public provision), we will continue to devote an enormous amount of time to litigation about inequitable, immoral and iniquitous gifted and special education programmes.

NOTES

1 Byrt and Gill (1973)

2 Raven et al (1988)

3 The conative components are those concerned with determination, persistence, and will. In the American literature other than that associated with McClelland, these components have either been ignored or inappropriately subsumed under 'affective'. Yet a person can very much enjoy doing something without being determined to see it through - and he or she can hate doing something, but still be determined to do it.

4 This does not mean that it is not useful to think about behaviour in terms of these categories. It only means that, in practice, attempts to assess the components separately are mistaken.

5 While it may be thought that the viewpoint developed here might be reconciled with traditional factor-analytic theory by focusing on qualities like 'the ability to make one's own observations', a little reflection shows that this is not the case. Our argument is precisely that such qualities cannot be assessed independently of valued goals. They have no *generalised* meaning. Therefore, they cannot be assessed by factorially-pure scales.

6 Raven, Johnstone and Varley (1985)

7 Raven (1984)

8 Bullock (1975)

9 Raven (1984)

10 Raven (1988)

11 Wolf (1987)

12 Ingenkamp (1977)

13 Hunter and Hunter (1984). The Hunter's main point is that intelligence tests are the *only* tests which account for a substantial proportion of the variance in occupational performance. Across thousands of studies of a wide range of occupational groups, they account for about 25% of the variance in performance. (The figure drops to 9% in longitudinal studies.) No other measures - of interests, biography, handwriting or anything else (including assessment center-based ratings) account for more than 1% of the variance. A good Vocabulary test which

takes only 5 minutes to administer will do as well as long more complicated, "intelligence" tests (Raven, Court and Rav 1988).

14 Berg (1973)

15 Hunter and Hunter (1984)

16 Berg (1973)

17 Folger and Nam (1964), O'Toole (1975)

18 Eysenck (1953)

19 Joint Committee for Educational Programmes (1981)

20 Flynn (1987)

21 e.g. Raven (1988)

22 See Raven (1977), Raven, Johnstone and Varley (1985), Rav (1981)

23 See Raven (1984), Raven (1985)

24 Raven (1988), Raven (1988), Raven (1987), Raven (1984), Rav (1984)

REFERENCES

Berg, I. (1973). Education and Jobs: The Great Training Robbery. London: Penguin Books.

Bullock Report. (1975). A Language for Life. London: HMSO.

Byrt, E. and Gill, P.E. (1973). Standardisation of Raven's Standard Progressive Matrices and Mill Hill Vocabulary for the Irish Population: ages 6-12. Cork, Ireland: MA. Thesis, National University of Ireland, University College Cork.

Eysenck, H.J. (1953). Uses and Abuses of Psychology. London: Penguin Books.

Flynn, J.R. (1987). Race and IQ: Jensen's Case Refuted. p221-232 in Modgil, S. & C. (eds). Arthur Jensen: Consensus and Controversy. Lewes, England: Falmer Press.

Folger, J.K. and Nam, C.B. (1964). Trends in Education in Relation to the Occupational Structure. Social. of Education. Fall. 19-34.

Hunter, J.E. and R.F. (1984). Validity and Utility of Alternative Predictors of Job Performance. Psychol. Bull. 96. 72-98.

Ingenkamp, K. (1977). Educational Assessment. Windsor, Berkshire: NFER.

Joint Committee on Standards for Educational Evaluation. (1981). Standards for Evaluations of Educational Programmes, Projects and Materials. New York: McGraw Hill Book Co.

O'Toole, J. (1975). Human Resources Development and Competency-Based Education. Syracuse University Research Corp. Educational Policy Research Center.

Raven, J. (1977). Education, Values and Society: The Objectives of Education and the Nature and Development of Competence. London: H K Lewis; New York: The Psychological Corporation.

Raven, J. (1980). The Most Important Problem in Education is to Come to Terms with Values. Oxford Review of Education, 7, 253-72.

Raven, J. (1984). Competence in Modern Society: Its Identification, Development and Release. London: H.K. Lewis.

Raven, J. (1984). Economic Policy in Modern Society. London: The Tawney Society.

Raven, J. (1984). The Role of the Psychologist in Formulating, Administering and Evaluating Policies Associated with Economic and Social Development in Western Society. J.Econ.Psychol., 5, 1-16.

Raven, J. (1984). Some Limitations of the *Standards*. Evaluation and Program Planning, 7, 363-370.

Raven, J. (1985). The Institutional Framework Required for, and Process of, Educational Evaluation: Some Lessons from Three Case Studies. In Searle, B. (ed). Evaluation in World Bank Education Projects: Lessons from Three Case Studies. Washington, DC: The World Bank, Education and Training Dept. Report EDT5 141-170.

Raven, J. (1987). The Role of the Psychologist in the Modern Economy. Proc. ESRC/BPS Conference on the Future of the Psychological Sciences, 122-140.

Raven, J. (1988). The Assessment of Competencies in Black, H.D. and Dockrell, W.B. (eds). New Developments in Education Assessment: British Journal of Educational Psychology, Monograph Series No. 3, 980126.

Raven, J. (1988). Choice in a Modern Economy: New Concepts of Democracy and Bureaucracy. In Maital, S. (ed). Applied Behavioural Economics. Brighton, England: Wheatsheaf.

Raven, J. (1988). Giving Information Teeth. Submitted to New Universities Quarterly.

Raven, J. (1988). A Model of Competence, Motivation and its Assessment. In Berlak, H. (ed). Assessing Academic Achievement: Issues and Problems. Madison Wisconsin: National Center for Effective Secondary Schools.

Raven, J., Johnstone, J. and Varley, T. (1985). Opening the Primary Classroom. Edinburgh: The Scottish Council for Research in Education.

Raven, J.C., Court, J.H. and Raven, J. (1988). A Manual for Raven's Progressive Matrices and Vocabulary Tests. London, England: H.K Lewis; San Antonio, Texas: The Psychological Corporation.

Spencer, E. (1979). Folio Assessments or External Examinations? Edinburgh: Scottish Secondary Schools Examinations Board.

Wolf, A. (1987). Work Based Learning: Trainee Assessment by Supervisors. Bradford, England: MSC Sales, ISCO.

RESUME

Les hypotheses suivantes, couramment faites par le psychometriciens et les utilisatuers de tests, sont remises e question:

- les scores des tests doivent etre distribues "normalement" - le tests doivent etre neutres par rapport aux valeurs - les tests doiver avoir une forte consistence interne - les psychometriciens ne doiver pas se preoccuper des convictions sociales ou politiques des sujets les personnes peuvent etre decrites independemment de la situatio dans laquelle elles se trouvent - nous avons le plus de chance d trouver un moyen economique d'aborder les differences individuelle en recherchant un petit nombre de variables resumant la majeur partie de la variance relative aux questions et aux aptitude humaines - il est important de trouver des moyens de differencie finement les personnes par rapport a un petit nombre d dimensions - la pratique consistant a rediger des rapport descriptifs des personnes n'est pas scientifique - proceder a de differenciations fines relatives a un petit nombre de dimensions es "objectif", par contre decrire les caracteristiques qui differencient le personnes ou relatives aux effets des programmes ou des politique educatifs et sociaux ne l'est pas - pour repondre a des objectif scientifiques ou pratiques tout trait defini doit etre stable.