Aalborg Universitet

Assessment and Educational Development

Graaff, Erik de

Publication date:
2000

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Graaff, E. D. (2000). *Assessment and Educational Development*. Aalborg Universitet. VCL-serien No. nr. 11

Assessment and educational development

by Erik de Graaff, Professor, Ph.D.

Assessment and educational development

VCL-serien nr. 11

ISSN	1399-7300
ISBN trykt udgave	87-90934-11-3
ISBN elektronisk udgave	87-90934-10-5

Udgivet af:
Videncenter for Læreprocesser Aalborg Universitet

Trykt af:
Vester Kopi as, Aalborg

VCL-serien redigeres af:
Susanne Nielsen, Brian Kjær Andreasen, Anette Kolmos, Annie Aarup Jensen og Palle Rasmussen

Distribution:
Videncenter for Læreprocesser
Att.: Susanne Nielsen
Fredrik Bajers Vej 7B
DK 9220 Aalborg Øst
Tlf. 9635 9955, Fax 9815 6542
E-mail: sn@vcl.auc.dk
Web-side: http://www.vcl.auc.dk

Assessment and

educational development

by Erik de Graaff
Professor, Ph.D.

VCL
June 2000

Assessment and educational development

E. de Graaff, Aalborg University, Videncenter for Læreprocesser, June 2000

Introduction

Assessment has always been an integral part of education. Usually, the teacher comments regularly on students performance during a course and at the end of a learning period there is some sort of an examination. Traditionally, such evaluations were based on questioning the students and observations of their performance by the teacher. By now we are aware that such judgements suffer from subjective bias (Van der Vleuten et al, 1991). The interpretation of criteria may vary from one teacher to the next. As a consequence of the subjective bias, comparisons between teachers, schools and educational systems are handicapped. This effect is clearly demonstrated by the first experiment on educational measurement on record, in Boston at the middle of the last century (Noll, 1961). The performance of 530 students on the same test was compared (a written examination with 154 questions, covering subjects like: arithmetic, geography, vocabulary, grammar and science). Results showed shocking differences between schools and teachers.

In this paper some highlights of educational measurement will be reviewed. Next, the relationship between the criteria of traditional educational measurement and educational innovations like problem-based learning will be discussed. The last paragraph describes the emerging of a new paradigm for assessment.

Educational measurement

In the trail of the Boston experiments the development of educational measurement as a discipline started around the turn of the century. Right from the beginning statistics were an important element. For instance, measurement characteristics, like reliability and validity of measures are expressed in terms of correlation with concurrent or criterion measures. Reliability is actually defined as the degree to which repeated measurements of the same trait produce the same result. According to this line of reasoning, the reliability of the test puts a limit to the validity: the maximum validity is the square of the test reliability. In this sense reliability precedes validity.

Subjective rater bias was recognized as one of the factors decreasing the reliability - and therefore also the validity - of educational measures. The answer to this problem was the “objective” test. During the first World War, the Army Alpha for classification of military personnel was the first objectively scrabble standardized test to be used at a large scale. The success resulted in numerous applications of the measurement techniques in education. Since World War II we can say that objective testing has really conquered the world. Billions and billions of students all over the world are tested each year with true-false or multiple choice tests.

The success story of objective testing is not undisputed. Right from the beginning both students and teachers seriously criticized these testing methods. Among the most heard complaints are: Ayou can guess the right answers”, Ait is just a gamble test” and Athose tests measure just recollection≅. Frequently, it is suggested that objective testing restricts study activities to the lower cognitive areas. The body of research on this issue is enormous and the results are ambiguous. Based on extensive literature search and several experiments Wesdorp et. al. (1979) conclude that empirical evidence for this claim is hard to find. According to these authors, the influence of testing on students' behavior depends on the status of the test. If the test results directly influence study progress or career opportunities, students will do anything to improve their performance. Routine tests are taken routinely and do not notably influence study behavior.

Differences in performance on different test formats (i.e. multiple-choice versus open-ended questions) is another major issue in research on educational measurement. Frederiksen (1984), for instance, cites a substantial number of studies suggesting that there appears to be a tendency to neglect higher cognitive skills, because of the extensive use of objective testing. However, such studies are methodologically difficult. A nice example is the carefully designed study by Newble et al (1979).The results of different groups were compared on a test cut in halves, where one group made the first half in the open-ended format, followed by a multiple choice half and the other group vice versa. Careful analysis of the data, however, revealed that the claimed difference in test format could be explained perfectly by the difference in difficulty between the two test halves (Galesloot et al (1982).

A further difficulty in comparing test formats, is that the difference is not so big as is looks. In fact only the automatic scoring procedure of objective tests is free from human influence. Whether the answer to a test question has to be chosen from a few alternatives or has to be written down in a short statement is just a matter of convenience. Therefore, Van der Vleuten et al (1991) coin the term Aobjectified≅ testing, indicating that objectivity is an attribute that tests possess to a certain degree. The point is not whether one test format is better than another. In some cases it is just more convenient to use objectified tests, in other cases different criteria prevail (see also: Norman et al, 1991).

The real issue is - or should be - to what effect testing is used and to what degree those objectives are realized. Basically, this is the question of the test validity: to what degree does a test measure what it is intended to measure? At this point it becomes clear that defining validity as being dependent of reliability results in a paradox. Measures need to be objectified in order to ensure that the conditions of repeated runs are as equal as possible. At the same time the conditions of objectified testing depart from the measurement objective. The psychologist A.D. de Groot called this phenomenon Athe problem of incomplete coverage". The operationalisation of an artificial construct (like intelligence) never completely covers the original construct-as-intended (de Groot, 1961). Construct validity reflects the degree of coverage of the construct-as-intended. With educational measurement, in any case but the straightforward testing of factual knowledge, the restrictions that are necessary to maximize reliability, decrease the construct validity.

Construct validity has to be estimated, it cannot be computed. Psychometricians prefer indices that can be calculated. For instance, a widely applied criterion for the validity of examinations is the prediction of future performance. Reviewing the literature on licensure examinations, Kane (1982) points out that there are so many factors influencing professional performance, that it is practically not possible to validate examinations this way. After some considerations he concludes that it is not even desirable to do so. He suggests to look for validity criteria within the context of the educational program: *AThe validity of a test .... would depend on how important the abilities being measured are for professional practice ....≅* (Kane, 1982). Actually, this implies that subjective judgement (what is important for professional practice) is being reinstated. In a similar fashion Hager et al (1994) discuss the introduction of ACompetency-based assessment≅ in education for the professions in Australia. Among the criticisms they treat are the lack of reliability and validity and the subjectivity of this assessment approach.

On the whole, the evidence concerning effects of testing on students' behavior is somewhat ambiguous. Partly this is caused by methodological difficulties. It is strongly suggested, however, that the dominance of statistical criteria has obscured the importance of the measurement objectives. Even if educational goals that are not adequately represented in examinations are not neglected by the students, at least they are not stimulated to follow a course towards those goals.

Educational measurement and the learning process

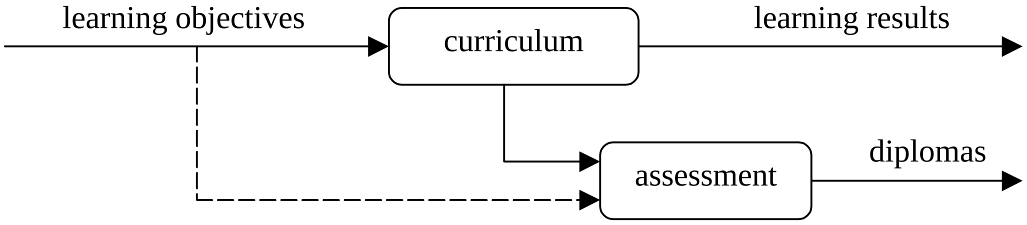
Originally, the wish to be able to do something lies at the bottom of all learning. As society got more complex, and education institutionalized, the distance between the wish of the learner and what happened in the classroom gradually became bigger and bigger. Learning in school is split up in different subjects and disciplines. Usually, the reasons for including those disciplines in a curriculum are still founded in their functionality for a profession. In many cases, however, students are unable to see that connection. In their perception, they just have to do what the teacher tells them to do. Consequently, the students learn for a grade, rather than to acquire a skill.

Re-installation of the link with professional practice is an essential aspect of the educational method Problem-based learning (PBL). In PBL, students typically work together in small groups, discussing practice related problems (Barrows & Tamblyn, 1980; Schmidt, 1983). In stead of being told what to do by a teacher, the problems appeal to the students' intrinsic motivation. Discussing problems activates prior knowledge relevant to the issue and fits new knowledge in a practice related cognitive structure (de Vries et al, 1989). Focussing on learning as a process implies that working on a problem is not just about finding the right solution. In learning to apply knowledge from an information base, the student exercises problem-solving skills. Key processes at the heart of the model for Problem-based learning developed by Pacific Crest Software are: Communication, Teamwork, Thinking, Use of Technology and Assessment (Apple et al, 1992). Learning in a PBL environment requires specific skills of the student, like chairing and participating in a meeting (Woods, 1994). Consequently, learning objectives in the area of cooperative skills become an integral part of education (Duncan-Hewitt et al, 1994). Students benefit both ways, once because their learning is motivated in an enhanced learning environment, and once because the team work skills come in useful when working in practice.

Integrating knowledge and skills from different disciplines may be fine from the perspective of the learner. It presents some problems, however, for assessment according to the traditional principles of educational measurement. It is much harder to define the learning objectives, or rather, the learning objectives become less homogenous. In a single subject course there is a limited set of test questions, whereas students select their own specific learning goals in a multi-disciplinary PBL course. As a consequence the reliability of measures (depending on the reproducibility of the measurement results) drops. Also, the added objectives, like communication and teamwork skills are hard to measure in an objectified manner. Assessment of such skills normally works with human judges, thereby increasing the subjective bias. In other words, it becomes difficult to meet the traditional criteria of educational measurement.

Within the framework of PBL this problem has been recognized. In the first years after the implementation of PBL in Maastricht it was noted that students tended to stick to their old habit of learning for an examination. At that time the assessment consisted of a block test at the end of an educational period (six each year). Instead of really exploring their own learning goals some students tried to find out what would be in the test, in order to maximize their examination results. In order to deal with this phenomenon, a completely new system of assessment was designed. In the traditional format the curriculum is constructed based on the learning objectives generated by the needs of society. Next, the assessment is derived from the curriculum. In the alternative approach, the assessment was derived directly from the learning objectives (see figure 1.).

Figure 1. The relationship between learning objectives and assessment



The assessment method, known as the Maastricht Progress Test is a large examination consisting of over 200 true-false questions (Verwijnen et al, 1982). This test aims at the end level of the curriculum and is taken by all students four times a year. The gain in knowledge takes the shape of a growth curve as the students progress through the curriculum. Since the test is developed separately from the curriculum, there is no way students can prepare for this examination, except to study as best as they can.

Still, the conflict between the objectives of innovative educational methods and educational measurement remains. As long as you operate within the assumptions of the rationalistic paradigm, you encounter these criteria for quality of measurement and the limitations they bring along. An example is the development of a method for the assessment of medical problem-solving abilities (de Graaff et al, 1987; de Graaff, 1989). The test called Simulation of Initial Medical Problem-solving (SIMP) was designed within a problem-based medical curriculum as an extension to the Maastricht Progress Test. Congruent with the educational principles, the SIMP-test is case based, focussing on application of knowledge in the context of practice. As a standard the responses of experienced physicians were employed. One of the difficulties that arose in constructing the test, was that the physicians often did not agree on a single right response. In fact they maintained that what was right or what was wrong depended largely on the context. Even when one physician recommended immediate hospitalization and another wanted to wait a few days, they claimed that both approaches could be qualified as responsible medical actions.

The solution to this problem was simple: all different right courses of action were included in the grading standard. From the traditional measurement perspective, however, this is abhorrent. The next paragraph will deal with this issue of measurement at a more fundamental level in describing the emerging of a new paradigm for assessment.

A new paradigm for assessment

Traditionally, assessment operates within the boundaries of a paradigm characterized as AScientific Measurement≅ (Hager & Butler, 1994). Because these terms suggest a claim on scientificness, Gubba and Lincoln (1983) prefer the term ARationalistic≅ paradigm. The foundation of a paradigm consists of a set of axioms, assumptions that can not be proved or disproved. The corner stone of the rational, or scientific measurement paradigm is static. In order to measure something it has to remain constant, at least for a reasonable period of time. In this view, a process is seen as a series of successive events, much like in a movie the illusion of movement is created by rapidly displaying a succession of pictures. Only within this static framework reliability could be defined in terms of repeated measurements and validity in terms of prediction of future performance. Objectivity is the second hallmark of the rational paradigm. Naturally, it makes sense - within this framework - to focus on the object being measured. If you want to predict performance on a fixed set of external criteria, you need absolute qualifications.

No matter how much it has dominated the scene, the rational paradigm is not the only one that can applied to educational measurement. The evaluation literature features a multitude of evaluation models, like: Naturalistic evaluation, Illuminative evaluation, Responsive evaluation, Adversary evaluation, Judicial evaluation (see: Walberg & Haertel, 1990). The list of different adjectives to be used in conjunction with evaluation is virtually endless. Van Berkel (1984) counts no less than 76 different types of evaluation. The choice of paradigm is obscured by the implicit dominance of the values of the rationalistic paradigm. For instance, the quality of assessment in the judgmental paradigm (Hager & Butler, 1994) is defined in terms of inter-subjective agreement among raters. Using this criterion, is at least paying lip service to objectivity. Table 1 summarizes the assumptions of the traditional scientific measurement paradigm with the process counterparts.

Table 1: Assessment Paradigms

Traditional assessment	**Process assessment**
static	*Dynamic*
- repeated measures	- unique moments
- prediction of future performance	- description of on going change
objective	*Subjective*
- absolute	- variable
- external criteria/responsibility	- internal criteria/responsibility

Starting with the first axiom, the focus of process education is dynamic rather than static. Learning is an ongoing process. Even if the learner encounters things for a second or third time, it is not exactly the same. Like the Greek philosopher Herakleitos, who stated you could never step in the same river twice. The natural point of reference for assessment is the here and now. The French phenomologist Merleau-Ponty points out that as the actual moment moves with time, our perspective of past and future changes. Consequently, subjective experience of the context becomes part of assessment. Also internal variable criteria replace absolute external criteria.

That it is possible to design a measurement procedure based on these principles is demonstrated by a personality test, called the Self-confrontation method (Hermans, 1974). The test consists of a set of questions, asking the respondent to reflect on a set of value areas, like Aa person that has been important in your past≅. Next, the person is asked to indicate the frequency of a set of 24 feelings in relation to the items on the personal list. The resulting measure is purely individual and may even vary over time with the same person. Comparisons within the value areas of one person or between persons are possible because the same set of feelings is used each time.

Discussion

The values implied by the core assumptions of the rational scientific measurement paradigm have dictated quality of educational measurement for almost a century. These values are clearly at odds with the objectives of process education. Problem-based learning challenges students to take responsibility for their own learning. As a consequence the learning activities vary from one student to the next. By definition, you can not predict what they are exactly going to do. They may even go outside or surpass the expertise of the teacher. This very fact often lies at the bottom of resistance against educational innovation. Traditional teachers love the average student. They despise the failures and stupid ones, but they really hate the brilliant students, the ones that outshine them and make their lives miserable by asking questions they cannot answer. A nice example of the way mediocre students are favored by traditional assessment is described by Pirsig in his bestseller of the seventies: “Zen and the Art of Motorcycle Maintenance”. As a teacher in a small Town, his leading character decides to drop the grading of assignments. First, his students are baffled. After a while the students who knew they could expect low grades and those who knew they were doing alright, relaxed. Free of judgement, their performance actually improved. Only, the students in the middle were not satisfied. If you know you are a C you want the hard work you put in and the little bit of luck you can get visibly rewarded with a B.

Pirsig continues his quest for quality in his second novel “Lila”. His main character feels that the woman Lila who travels with him on a boat for a while possesses quality. He tries hard to define that quality against the background of her obvious failure to manage her life. Since he is a sincere and consequent thinker, the outcome is inevitable. Just like More and Groce contemplating the concepts Agood” and Abeauty≅ respectively, the result of hundreds of pages of arguments is that quality can not be defined outside of the context.

The challenge for educational measurement is to design instruments and procedures that can be used within a dynamic and subjective framework. Also the methodology must be developed to validate these assessment tools against criteria that are consistent with the process paradigm. Consequently, criteria like reliability and validity must be re-defined. Instead of repetition, reliability could for instance be defined in terms of trustworthiness, emphasizing that the students’ own judgement should be an integral part of assessment. Validity could be re-defined in terms of the degree to which assessment contributes to improvement. Applied to assessment itself, that would mean that assessment is good as long as it contributes to the general objectives of education.

References

Apple, D., Beyerlein, S.W. & Schlesinger, M. A. (1992) **Learning Through Problem-solving.** Corvallis, OR: Pacifis Crest Software Inc.

Barrows, H.S. & Tamblyn, R.M. (1980) **Problem-based learning, an approach to medical education.** New York: Springer.

Berkel, H.J.M. van (1984) **De diagnose van toetsvragen.** [Diagnosis of test questions] (Ph.D. thesis) Amsterdam: Centrum voor Onderzoek van het Wetenschappelijk Onderwijs.

Frederiksen, N. (1984) The real test bias; Influences of testing on teaching and learning. **American Psychologist**, 3, 193-202.

Duncan-Hewitt, W., Mount, D..L., Apple, D.(1995) **A Handbook on Cooperative Learning** (sec. ed.) Corvallis, OR: Pacifis Crest Software Inc.

Galesloot, J.A.M., Graaff, E. de, Verwijnen, M. and Imbos, T. (1981) Facts and beliefs on the issue multiple-choice testing vs. free-response tests in examinations of clinical competence. **Medical Education**, 15, 204-205.

Graaff, E. de, Post, G.J. and Drop, M.J. (1987) Validation of a new measure of clinical problem-solving. **Medical Education**, 21, 213-218.

Graaff, E. de (1989) **Simulation of Initial Medical Problem-solving: Studies on a new measure for the assessment of medical problem-solving ability.** (Ph.D. thesis Rijksuniversiteit Limburg, Maastricht). Haarlem: Thesis.

Gubba, E.G & Lincoln, Y.S (1983) In: George F. Madaus, Michael Scrive & Daniel L. Stufflebeam (eds.) **Evaluation Models; Viewpoints on educational and human services evaluation.** Boston, the Hague: Kluwer - Nijhoff Publishing

Hager, P. & Butler, J. (1994) Problem-based Learning and paradigms of assessment. In: S.E. Chen, R.M. Cowdroy, A.J. Kingsland and M.J. Ostwald (eds.) **Reflections on Problem-based Learning.** Sydney: Australian Problem Based Learning Network.

Hager. P., Gonczi, A. & Athanasou, J. (1994) General Issues about Assessment of Competence. **Assessment & Evaluation in Higher Education**, Vol. 19. No. 1. 3-16.

Hermans, H.J.M. (1974) **Waardegebieden en hun ontwikkeling.** [The development of value areas] Amsterdam: Swets en Zeitlinger.

Kane, M.T. (1982) The Validity of Licensure Examinations. **American Psychologist**, 37, 911-918.

Newble, D.I., Baxter, A. & Elmslie, R.G. (1979) A comparison of multiple choice testes and free response tests in examinations of clinical competence. **Medical Education** 13, 263-268.

Norman, G.R., Vleuten, C.P.M. van der & Graaff, E. de (1991) Pitfalls in the Pursuit of Objectivity: Issues of Validity, Efficiency and Acceptancy. **Medical Education**, 25, 119-126.

Pirsig, R.M. (1974) **Zen and the Art of Motorcycle Maintenance.** New York: Bantam books.

Pirsig, R.M. (1991) **Lila; An inquiry into mortals.** New York: Bantam press.

Schmidt, H.J. (1983) Problem-based learning: rationale and description. **Medical Education,** 17, 11-16.

Vleuten, C.P.M. van der, Norman, G.R. & Graaff, E. de (1991) Pitfalls in the Pursuit of Objectivity: Issues of Reliability. **Medical Education**, 25, 110-118.

Verwijnen, G.M., Imbos, T., Snellen, H., Stalenhoef, B., Pollemans. M., Van Luyk, S., Sprooten, M., Van Leeuwen, Y. and Van der Vleuten, C.P.M. (1982) The evaluation system at the medical school of Maastricht. **Assessment and Evaluation in Higher Education**, 7, 3, 225-244.

Vries, M.W de., Schmidt, H.G. and Graaff, E. de (1989) Dutch comparisons: the assessment of cognitive and motivational effects of problem-based learning. in: H.G. Schmidt, M. Lipkin, M.W. DeVries and J.M. Greep (eds) **Education for tomorrow's doctors today.** New York: Springer-Verlag.

Walberg, H.J. & Haertel, G.D. (eds) (1990) The International Encyclopedia of Educational Evaluation. Oxford: Pergamon press.

Wesdorp, H. (red.) in samenwerking met Blok, H., De Graaff, E., Wolowitsj-Schelvis, A. and Zijlmans, S. (1979) **Studietoetsen en hun effecten op het onderwijs.** SVO-reeks, 15, 's-Gravenhage: Staatsuitgeverij.

Woods, D. (1994) **How to gain most from problem based learning.** Hamilton, Ontario: Mcmaster University.

I VCL-serien er udgivet følgende numre:

Rittenhofer, Iris. (1999) Askepot bager luksuskringle: kønsbarrierer i de højere uddannelser og i forskningen. Arbejdspapir VCL. Aalborg: Videncenter for læreprocesser, Aalborg Universitet. (VCL-serien nr. 1 1399-7300)

Kjær Andreasen, Brian ; Kolmos, Anette. (1999) Undervisningsportfolios på højere uddannelses-institutioner. Pædagogisk Udviklingscenter og Videncenter for læreprocesser, Aalborg Universitet.(VCL-serien nr. 2)

Algreen-Ussing, Helle ; Keiding, Tina Bering ; Kolmos, Anette. (1999).Pædagogisk omstilling, læringsopfattelser og organisatoriske rammer. Aalborg: Pædagogisk Udviklingscenter og Videncenter for læreprocesser , Aalborg Universitet.(VCL-serien nr. 3)

Jacobsen, Lone (1999). Internationalisering i ungdomsuddannleserne - resultater fra en pilotundersøgelse. Aalborg: Videncenter for læreprocesser, Aalborg Universitet (VCL-serien nr. 4)

Lorentsen, Annette (2000). Aspekter af teknologistøttet fjernundervisning på universitetsniveau. Aalborg: Pædagogisk Udviklingscenter og Videncenter for læreprocesser, Aalborg Universitet. (VCL-serien nr. 5)

Kloch Frederiksen, Birte (2000). Den verdensfjerne videnskab? En diskursanalyse af Ingeniørens Ugeblad 1970-1974. Aalborg: Videncenter for læreprocesser, Aalborg Universitet. (VCL-serien nr. 6)

Aarup Jensen, Annie (red.) (2000). Fornyelse af egen praksis - eksperiment og refleksion i sprogundervisningen. Aalborg: Videncenter for læreprocesser, Aalborg Universitet. (VCL-serien nr. 7)

Kolmos, Anette (red.) (2000). Online Læring - lærerkvalificering, didaktik og kommunikation. Aalborg: Pædagogisk Udviklingscenter og Videncenter for læreprocesser, Aalborg Universitet. (VCL-serien nr. 8)

Kloch Frederiksen, Birte (2000). Moral, solidaritet og fællesskab. Ph.d.-forelæsning: med efterkrift af direktør, professor Palle Rasmussen. Aalborg: Videncenter for læreprocesser, Aalborg Universitet. (VCL-serien nr. 9)

Langeland Christensen, Jonna & Storgaard, Stine (2000). Forsøg med undervisningen i projektskrivning- og arbejdsprocesmetoder på Den Samfundsvidenskabelige Basisuddannelse. Aalborg: Pædagogisk Udviklingscenter og Videncenter for læreprocesser, Aalborg Universitet. (VCL-serien nr. 10)

Graaff, Erik de. (2000) Assessment and educational development. Aalborg: Videncenter for Læreprocesser, Aalborg Universitet. (VCL-serien nr. 11)

Langeland Christensen, Jonna. (2000). Evaluering af forsøg med undervisningen i projekt- og gruppearbejde på Den Teknisk-Naturvidenskabelige Basisuddannelse. Aalborg: Pædagogisk Udviklingscenter og Videncenter for Læreprocesser, Aalborg Universitet. (VCL-serien nr. 12)