European Conference on Educational Research

BOOK OF SUMMARIES

Volume 1

University of Twente
The Netherlands
June 22 - 25, 1992

COLOFON

Editors:
Tjeerd Plomp
Jules Pieters
Andries Feteris

Editorial assistance:
Harmen Abma
Jeroen Breman
Conny de Koning
Olivia Kramers
Renate Schraa

Cover:
Hanna Snijder

Print:
Duoprint

University of Twente
Department of Education
P.O. Box 217
7500 AE Enschede

ISBN 90-365-0534-8

SISS-data were collected in 1984 and relate to pupils with a mean age of 15 years and six months. In SIMS achievement was measured by means of a multiple choice mathematics test. The multiple choice test used in SISS consists of items about physics, chemistry, biology and earth science.
In the analyses the following variables served as covariates:

- Sex;
- Social economic status;
- Achievement motivation;
- Cognitive aptitude;
- Type of education;

It was also investigated if interaction-effects could be discerned between school size and one of these first four covariates. If interaction-effect would be revealed this would imply that the effects of school size on student achievement differ for certain groups of pupils. In the analyses school size was operationalised as a categorical variable. Thus it would also be possible to reveal non-linear relations between school size and pupil achievement.
None of the analyses revealed a statistically significant effect of school size on pupil achievement. Of the several dozens of interaction-terms which were examined only one showed a statistically significant effect on pupil achievement: In the Netherlands the girls in schools with at least 360 but less than 500 pupils got better results on the mathematics test than there male classmates.

STANDARDIZED ACHIEVEMENT TESTING IN DIFFERENT CURRICULAR SETTINGS

W.J. Pelgrum, Department of Education, University of Twente, Enschede, The Netherlands; D.M. de Haan, Open University, OTIC, Heerlen, The Netherlands

Introduction

Although the past decade has shown in many countries an increased interest in monitoring educational progress by comparing achievement measures over time and between nations, methodological sophistication has lagged behind to warrant conclusions commonly based on these comparisons. Especially the interpretation of scores on standardized achievement tests in terms of educational effectiveness does not take into account curricular variation that may lead to differences in overlap of the test and curriculum for groups of students which are compared. This paper examines empirical evidence related to this problem, the nature and validity of measures for registering test-curriculum overlap, and it discusses results from a study for improving these measures.

Analyses of "old" measures

It has been argued that a number of basic conditions are essential for adequate educational assessment. Educational assessment is almost by definition a large-scale enterprise, involving large samples of schools, teachers and students. Given

the fact that many actors are involved in realizing and appraising the output of an educational system, from a communicative point of view it is essential that outputs of the main sub-systems are registered. A minimum requirement would be the measurement of the intended, implemented and realized curriculum as indicators of the output at macro-, meso- and micro-level of the system. An important problem to be solved is the adequate measurement of these different outputs. In many large-scale studies conducted by International Association for the Evaluation of Educational Achievement (IEA) a choice is made for item-based measurement of the different outputs as a basis for processing and evaluation of the measures of the realized curriculum. Pelgrum (1990) addresses the question of how valid and how applicable item-based measures of the implemented curriculum are (as used in the IEA studies). In order to shed light on the validity issue, secondary analyses of data collected in the Second International Mathematics Study (SIMS) and the Second International Science Study (SISS), conducted by IEA, were undertaken. The implementation measures collected in SIMS consisted of teacher and student ratings of test items on the question of whether the corresponding subject matter was taught, whereas for the Netherlands some additional measures of the same variable were collected. In the Dutch part of SISS the same teacher ratings as in SIMS were used. In addition, data for one country in SISS were available in which we use a multidimensional item-based rating of implementation by teachers. The results presented showed that in general the validity of the ratings by teachers is promising. The ratings of teachers corresponded with the content of the textbooks they used, while factor analyses of the ratings reproduced the (curriculum) structure of the Dutch school system. The differences between Dutch school types in subjects included in the lesson tables was reflected in the ratings. Comprehensive school systems showed much less differences in implementation than non comprehensive systems. However, it was also noted that teachers tended to under-rate the amount of subject matter presented to students. Furthermore, it was shown that for particular subsets of items teachers in the Netherlands made serious mistakes in Judging whether the corresponding subject matter had been taught before the date of testing. Pelgrum (1990) concluded that continuing work is needed to study the feasibility of item-based measures of the implemented and intended curriculum.

Analyses of "new" measures

De Haan (1992) addressed the question how ~he existing IEA measures might be improved. She constructed a revised version of the item-based measures as used in the IEA studies by not only asking teachers to judge for each item the difficulty and whether the corresponding subjectmatter was taught. but she also added ratings regarding the suitability of the terminology and the format of the item, such as terminology, format or symbols. She called this instrument D-TCO (Detailed Test Curriculum Overlap). Furthermore, she used a so called H-TCO (Holistic Test Curriculum Overlap) instrument asking teachers to select items they judged fair to administer to their students. In a pre- post test design teacher, textbook, and student ratings were collected. The results of this study are summarized below. With regard to the *reliability of* TCOjudgements collected with the D-TCO instrument It can be concluded that the stability of these ratings is acceptable if the Judgements of teachers are recoded to a dichotomy of whether or not an Item is

taught before the date of testing (D-TCO Judgements). With the D-TCO instrument, 95% of the items that were judged as taught at the pretest, were also judged as taught at the posttest. With the H-TCO instrument this percentage was 92%. Hence the stability of the H-TCO Judgements seems also acceptable. ~he stability of the unrecoded D-TCO judgements (that is, the judgements of whether an item is taught in a specific time period) is considerably lower. Judgements of whether an item was taught in elementary school are found to have the lowest reliability. With regard to the *construct validity of* the D-TCO instrument as well as of the H-TCO instrument, it can be concluded that on an aggregated level (for groups of teachers and items), the construct validity is reasonable. At the pretest as well as at the posttest, the D-TCO judgements and the H-TCO Judgements correlate in general the highest with each other (varying between r-.78 and r=.91) and correlate lower, but still significantly, with the textbook based TCO judgements (varying between r=.50 and r=.76) and with the student based TCO judgements (varying between r=.56 and r=.84). A comparison of the absolute differences between percentages of TCO based on different approaches, showed that the percentage of DTCO judgements is significantly different from the textbook based TCO judgements. It was supposed that by adapting the results of a textbook analysis for each individual teacher to differences in textbook use (as measured with the teacher questionnaire and with the registration forms filled in by teachers during the period between pretest and posttest), the textbook approach could be an appropriate measure of the operational curriculum. A possible explanation for the discrepancies between D-TCO and textbook based TCO is that the textbook based approach, although adapted to individual differences~ between teachers, is still predominantly referring to the formal curriculum, while the D-TCO measure reflects the operational curriculum. It was also shown that the different TCO measures were significantly higher at the posttest than at the pretest. Hence it seems to be possible to detect changes in a curriculum over time by use of TCO measures. No significant differences are found between D-TCO-measures of different school types. An explanation can be found in the selection of items: the items that were judged by teachers of both school types did not differentiate enough between the curriculum of both school types. The examination of the construct validity at a more specific level showed that the D-TCO instrument as well as the H-TC~O instrument are less convergent with the other approaches: analysing the convergence of different approaches for each teacher individually showed that DTCO judgements were convergent with H-TCO Judgements and with the textbook based TCO judgements. But the average percentage of convergent judgements per item of both the D-TCO and the H-TCO measure varied between 75% (sd=17) and 84% (sd=ll). The size of the standard deviations are quite high, which means that there is a great variation in convergency over items. This provokes the idea that the validity of Judgements varies over different Items. At this specific level, student-based rating appeared to be less convergent with the other approaches. With regard to the Judgement of whether an item deviates on one or more of the characteristics, it can be concluded that in general D-TCO Judgements are convergent with textbook based TCO judgements. However, if one only looks at the convergence of Judgements of items that are Judged as deviating on a specific characteristic, it was shown that D-TCO judgements are not convergent with textbook based Judgements. With regard to

the *predictive validity* of the D-TCO measure, it can be concluded that student outcomes are to some extent related to D-TCO judgements of whether an item is taught. However since the size of the correlations varies between .37 and .46 at the test level (that is measured as the correlation between average student test score per teacher and percentage of ~taught~ items per teacher), and between .37 and .50 (in Vocational Education) and between .32 and .45 (in General Secondary Education) at the item level (that is, the correlation between average student score per item and percentage of teachers judging an item as 'taught'), it might be questioned whether the D-TCO measure is a good predictor for student achievement. A multiple regression analysis in which beside the TCO rating, the judgement of deviation of an item on certain characteristics was used as an extra independent variable did not improve the prediction of student achievement. Comparing the correlations we found with the correlation of the IEA-TCO measure with student scores collected in SIMS (test level: r=.22, (n=229); item level: Domestic Science Education: r=.32 (n=40), General Secondary Education: r= .00, n=40)), reported by Pelgrum (1990), showed that the correlations of the DTCO measure are higher at the test level and for General Secondary Education at the item level, but that these differences are not statistically significant. An analysis of the predictive validity of the H-TCO judgements showed that the correlations between these judgements and student achievement were higher than the correlations between student outcomes and D-TCO judgements. At a more specific level, it was found that, for those items that were judged as taught between the pretest and the posttest (according to the D-TCO measure), the increase of student scores was significantly higher than for items that were not taught between the pretest and the posttest, even when controlled for differences in student characteristics. An analysis of the influence of perceived difficulty of an item (measured by teacher estimations of percentage correct) showed that at the item level the D-TCO judgements as well as the H-TCO judgements are strongly related to the perceived difficulty. With regard to the *efficiency* of the D-TCO and the H-TCO instrument, it was shown that the time needed to judge one item with the D-TCO instrument is about 4 times longer than with the H-TCO instrument. The mean time needed to judge an item with the D-TCO instrument was 87 seconds compared to ~3 seconds for the H-TCO instrument.

References

Haan, D. M., de, (1992). Measuring test-curriculum overlap. Enschede: University of Twente.

Pelgrum, W. J. (1990). Educational Assessment: monitoring, evaluation and the curriculum. De Lier: Academisch Boeken Centrum.