ED 472 706 IR 021 648

AUTHOR Mikk, Jaan
TITLE Experimental Evaluation of Textbooks and Multimedia.
PUB DATE 2002-00-00
NOTE 21p.; In: Selander, Staffan, Ed.; Tholey, Marita, Ed.; Lorentzen, Svein, Ed. New Educational Media and Textbooks: The 2nd IARTEM Volume. Stockholm Library of Curriculum Studies 9. Stockholm Institute of Education Press, 2002, p121-40.
PUB TYPE Reports - Descriptive (141) -- Reports - Research (143)
EDRS PRICE EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Efficiency; Foreign Countries; Junior High Schools; Multimedia Materials; Secondary School Science; Text Structure; *Textbook Evaluation; *Textbook Research; *Textbooks
IDENTIFIERS Estonia

ABSTRACT

This paper begins by providing an overview of three types of textbook research methods: asking teachers, parents, or students about the different aspects of textbook quality; textbook analysis, i.e. counting some characteristics of a textbook using strictly fixed rules; and the experimental evaluation of textbooks, usually carried out in schools. Determining the appropriateness of a textbook for Estonian students is discussed, including three issues: which students should take part in the experiment; which tasks should be composed to the content of the textbook; and which level of correct answers is the optimal one. An experimental comparison of the quality of two Estonian textbooks (a textbook of physics for grade 7 and a textbook of anatomy for grade 8) is then presented, based on indices of the efficiency of the textbooks. The problem of equalizing the condition of using the two textbooks is discussed. Summary results present findings for four indicators of efficiency-comprehension, acquisition, information gain, and persistence of knowledge. Unexpected results of the experimental research are described, including the validity of some text characteristics in predicting reading outcomes. (Contains 40 references.) (MES)

Jaan Mikk

Experimental evaluation of textbooks and multimedia

1

According to the bylaws of IARTEM the main purpose of our association is "to promote research on and understanding of textbooks and educational media." The research has been promoted at our conferences by representing the results of different interesting investigations. The results have deserved keen attention but usually there was little time to analyse the methods on which the results were based. I think that textbook research methods need special consideration by our association. The history of science gives many examples of the importance of developing research methods. If there is no new research method, the branch of science is in the danger of decadence. New research methods usually lead to new discoveries and florishing of the field of science.

Textbook research methods can be divided into three groups. Let us look at them very briefly.

Nowadays the most common method is to ask teachers, parents, or students about the different aspects of textbook quality. Very many questionnaires have been composed to obtain the assessments (Die Schulbuchbegutachtung ... 1991; Tyson-Bernstein 1989; Tholey 1996; Rauch & Tomaschewsky 1986; Vassilchenko 1995). The theoretical aspects of the expertise have been studied by V. S. Cherepanov (1991). The method is easy to implement and questions can be put to all the aspects of textbook quality. On the other hand, different experts may differently evaluate one textbook and therefore the evaluations are sometimes of questionable value.

The second group of methods is textbook analysis. The analysis consists in counting some characteristics of a textbook using strictly fixed rules. For example, counting the word length and the sentence length enables the researcher to calculate the readability index of a textbook. The analysis can be often computerised and carried out before using the textbook in school or even before printing. In the latter case, the unsatisfying text can be rewritten before printing the textbook. On the other hand, it is difficult to define the exact rules for counting all the important characteristics of textbooks and sometimes it is not clear if the data collected are really implementable

The experimental evaluation of textbooks is usually carried out in schools. The results of an experiment are the most reliable indicator of textbook efficiency and the results serve as basis for validating other methods of textbook evaluation. Therefore, the experimental evaluation is crucial in textbook research. On the other hand, experimental investigations take much time and need considerable funding. Experiments should not be carried out before the textbook analysis has revealed that there are no serious shortcomings in the textbook. Otherwise the development of the participants in experiment can be hindered.

The experimental investigation of a textbook can have two aims: we may want

1) to ascertain if the textbook under study could be used in school,

2) to ascertain which of the two or more textbooks is better.

We will discuss the two types of an experiment separately.

Determining the appropriateness of a textbook

One of the biggest problems in Estonian schools is the difficulty of the textbooks for children. The textbooks overload pupils, hinder the development of thinking and self confidence, demotivate them to read all the life. On the other hand, textbooks should not be too easy for pupils. How to find out which textbook is appropriate for pupils?

To solve the problem three issues should be considered:

1. which students should take part in the experiment,
2. which tasks should be composed to the content of the textbook,
3. which level of correct answers is the optimal one.

The first issue is simple to handle if the number of students in our interest group is small. In this case, all the students can take part in the experiment. However, in most cases the researcher is interested in the appropriateness of a textbook for a large number of students. *The students participating in the experiment should be representative to all the students in this case.* A representative sample can be formed by random sampling of students from the whole sample of potential users of the textbook. In practice, random sampling is seldom used as it is difficult to organise an experiment when there are only one or two students participating from a school. Therefore the whole sample is divided into subgroups and from each subgroup a specified percentage of students is invited to participate in the experiment. For example, if the whole sample has 40 percent of students studying in countryside schools, then the representative sample must also have 40 percent of participants from countryside schools. We see that the representativity of students in educational experiments is analogous to the representativeness of respondents in sociological investigations.

Which is the optimal size of a representative sample of students for the experiment to evaluate a textbook? The size of the sample depends on the desired exactness of results. The more students are involved in the study, the more precise the results will be. The number of students participating in an experiment does not depend on the size of the total student population, it depends on the diversity of the population and the allowed error of measurement as can be seen in the Formula (1).

$$n=\frac{t^2\delta^2}{(\Delta x)^2}$$

(1)

where
n - number of students in experiment,
t - Student's coefficient,
_ - standard deviation of results,
_ x - allowed error of measurement.

To use the formula, we have to know the approximate value of the standard deviation of the results and we have to fix the allowed error of measurement. The approximate value of standard deviation can be estimated in preliminary experiments and the Student's coefficient can be found in statistical tables. Then the number of participants in the experiment can be calculated.

The second issue deserves more thorough analysis. The starting point of the analysis is the idea of representativness of the test that can be composed considering the content of the textbook. The methods for *composing the representative set of test items* are the same as in composing a representative sample of students. In principle, the test items can be randomly selected from all the possible items in the textbook but the method is not used because an all-covering set of items is usually not available. Therefore it is important to classify all the elements of the textbook content and to compose test items so that the numbers of the items in all the classes are proportional to the element numbers of these classes of the textbook.

There are many possibilities to classify the elements of the textbook content. For example, if 30 percent of information is given on illustrations in the textbook then there should be composed 30 percent of test items on the content of the illustrations. The other basis for the classification of textbook elements might be the level of acquisition according to Bloom, the grouping of the content items and others.

The number of items in a textbook referenced test is a problematic issue. In principle, the number of items should be determined as the representative sample of students was suggested to be obtained, i.e. according to Formula (1). It means that several hundred items are needed to evaluate a textbook. Experimental investigations have proved that about 400 items are needed to obtain the results with an error of measurement lower than 5 percent of their extent in 95 percent of cases (Mikk 1981, 93).

Let us consider another aspect of composing questions to a textbook. It is well known that some questions for a text may be difficult and the other questions for the same text may be easy. How can we assess textual difficulty if the answers to the questions depend heavily on the characteristics of the question? To assess the text, the questions should have the same level of difficulty as the text.

J. S. Chall (1958, 40) writes about an investigation in which a correlation 0.62 was found between the complicacy of texts and the complicacy of questions formulated for the texts. It seems to be a relatively high correlation but nevertheless the a correlation is 0.78. Consequently about 60 percent of the variation in the complexity of questions did not correlate with the complexity of texts. These questions do not enable the researcher to assess the difficulty of texts.

Analogous results were later achieved by E. B. Entin and G. R. Klare (1980). They found that some multiple-choice questions were answered correctly by 80 percent of testees without reading the corresponding text. It distorts the difficulty indices developed on answers to multiple-choice questions.

We carried out an experiment to establish which characteristics of questions are correlated with their difficulty. We took 4 sections (about 500 words each) from a physics textbook, constructed 8 versions of questions for every text and 304 tenth grade students answered the questions after reading the passages. Every student answered only one version of questions on a text. There were altogether 320 questions under study.

We tried to compose equal test versions for a text. Nevertheless, the results of the experiment indicate that the versions from 10 questions were different. For example, the testees answered correctly 48 percent of questions in one version and 67 percent of questions of another test version on the same text. Of course, a part of the difference in these figures may be explained by the differences in students' abilities, who answered the versions but the influence of the difficulty of questions must be considered as well. To specify this influence, we calculated the correlation coefficients between the characteristics of questions and the percentage of correct answers to the questions. The correlations elicited the following factors of the difficulty of questions.

1. The questions on terms were answered better than the questions on facts or notions.
2. Fewer correct answers were given to the questions which had longer answers in the text.
3. It is more difficult to produce the correct answer when the number of concepts associated with the answer is larger.
4. The longer the words of the question, the fewer correct answers were given.
5. The percentage of correct answers was 71 for the questions based on the recognition of the material and 56 for the questions aimed at the reproduction of the material.

Analogous results were obtained by K. Green (1984). She varied multiple-choice answers to a test item and, as a consequence, the percentage of correct answers changed from 22 percent to 70 percent. Careful compilation of questions on a text is crucial in obtaining valid indices of its difficulty.

Multiple-choice questions are frequently used in tests. One of the choices is correct and the others are not. Sometimes the testees do not know the correct answer but, nevertheless, they mark one of the choices as correct. *Guessing of answers* is sometimes successful and so the testees get a somewhat higher result than their actual level of knowledge allows. This distorts the results. To calculate the actual level of knowledge, Formula (2.4) can be used.

$$A=\frac{R-\frac{W}{k-1}}{n}100\%$$

(2)
where:
A - achievement level of the testee in percentage,
R - number of correct answers,
W - number of incorrect answers,
k - number of multiple-choice answers to a question,
n - number of questions.

The formula is often used in scoring standardised tests. It is expressed in the rules such as subtract 0.25 points from the number of correct answers for every incorrect answer (Taking the SAT I... 1994, 76). The correction from the Formula (2) will be greater when the number of multiple-choice answers to a question is small. In the case of two alternatives the number of incorrect answers should be subtracted from the number of correct answers. Without using the Formula (2) we cannot figure out the actual results of textbook referenced tests.

There are also other methods for the assessment of the difficulty level of textbooks besides the answering of questions. *The cloze procedure* is the most promising. The method lies in deleting every n-th word in a text under study and in filling in the blanks by students. The higher the percentage of correct fulfilment, the easier is the text to be understood.

The cloze procedure seems to be very different from answering questions but actually the methods are similar. To produce a question, the investigator often deletes a word or a phrase in a sentence, substitutes the deleted word or words by a question word and rearranges the words according to the rules of interrogative sentences.

In some aspects the cloze procedure is a better method for measuring text difficulty than questions are. To produce questions, the investigator can substitute no matter which word with a question word and, therefore, the difficulty of questions depends on the investigator's choice. Contrary to this discretion of an investigator, in cloze procedure the word is deleted only by strict rules. Therefore the results of a cloze test should give exact indices for the comparison of difficulty of texts.

The comparison of questions and the cloze procedure also indicates some of the shortcomings of the cloze procedure. Always only one word is deleted to produce a blank in a cloze test but to produce questions sometimes a phrase is substituted by a question word. Questions on a text may be composed relying on two or more sentences but this possibility cannot be used in cloze procedures. Due to these shortcomings in the cloze procedure, its validity may be lower than the validity of questions especially in measuring comprehension on the inter-sentence level.

Nevertheless, the cloze procedure is an appealing method for measuring text comprehensibility. Many researchers have concluded that the cloze procedure gives better results than readability formulae (Hater & Kane 1970; Potter 1968; Weintraub 1968). J. R. Bormuth has written a survey about cloze procedure and found correlations 0.73 - 0.95 between cloze tests and answering questions (Bormuth 1968). I. A. Rapoport and his co-workers (1976) have received a correlation 0.96 between the integral indices of foreign language knowledge and the results of a cloze test. The indices of cloze test validity are relatively high to approve the use of the method for the measurement of text difficulty.

Let us proceed to the third issue of determining the appropriateness of a textbook for students: *which difficulty level is the optimal one*. It is obvious that too difficult or too easy a textbook is not the best. There is some optimal level of correct answers to the questions or correct fillings in the blanks of cloze procedure. The following overview will consist of three parts: the optimal level of text comprehension, text acquisition and cloze tests.

Text comprehension and text acquisition are measured by giving testees questions to answer. In both cases, testees should have enough time to answer all questions. The difference between the measurement procedures is the following. In measuring text comprehension, testees can use texts all the time for the formulation of their answers. In measuring text acquisition, testees study the text independently, then put the text away and answer the questions.

There is a standard for comprehension tests widely used in the USA. According to the standard, a text is suitable for independent study when the student can comprehend 90 percent of its content. A text can be studied with teachers' help when the student can independently answer correctly 75 percent of the questions set on the content of the text (Bormuth 1968).

The criteria are supported by the tradition of programmed learning. Many authors (Agur, Toim & Unt 1967, 95; Nikandrov 1970, 39) write that linear programs are suitable for students if they give 90-95 percent correct answers. Questions in linear programs are answered by using texts, therefore the criteria can be seen as criteria for text comprehension.

The specialists on reading H. P. Smith and E. V. Dechant (1961, 243-248) are convinced that a book is too difficult for children when they can understand less than 85 percent of its content. Obviously, it is the lowest comprehension level where the text can be used for independent study. When the comprehension of the text is 75-90 percent, it can be studied during supervised instruction.

The acquisition of study material is the most frequently used aim of education, therefore many specialists have written about its required level.

J. K. Babanskii (1977, 59) has claimed that the study material is reasonably well acquired when students can answer correctly at least 70 percent of questions. At the lower level of comprehension, the acquisition is not stable and students waste their time.

Specialists on programmed learning have given their students tests after learning a chapter. They allowed their students to go to next chapter when students answered 70 percent of test items correctly (Talyzina 1975, 306; Taranov 1976, 94-95).

Reading specialists also agree with the criteria. They write that acquiring 70 percent of text content is satisfactory (Kuznetsov & Khromov 1977, 30; Maanso 1969).

Theorists on mastery learning have studied the optimal level of acquisition. They are convinced that acquisition at the level of 80-90 percent is the most appropriate (Anderson & Block 1985). N. O. Cristoffersson (1971, 130-131) has studied the time needed for learning. He concludes that learning is most economical when the average level of acquisition is 80 percent. Then the able pupils acquire 100 percent and less able 60 percent of the study material. In our experiments with seventh grade students we have found that the average optimal level of acquisition should be 70 percent of study material in mathematics and history (Mikk 1981, 312).

The criteria of optimal values for cloze tests were studied by J. R. Bormuth. He found in a study that comprehension at a 75 percent level is comparable to 44 percent on a cloze test and 90 percent comprehension level is comparable to 57 percent in a cloze test drawn from the same passage (Bormuth 1968). In another detailed experimental study he found the following criteria of optimality: willingness to study was the highest with 50 percent correct answers on cloze test, difficulty preference ratings were the highest with 55 percent of cloze score, style preference ratings and subject matter preference ratings were the best with 70 percent of cloze score, the rate of reading was the highest with 72 percent of cloze score, and information gain was the largest by 80 percent of correct answers to cloze test (Bormuth 1971, 113). J. R. Bormuth has also elaborated summative optimal values of cloze test for grades 3 to 12. He has found, for example, that the optimal cloze score is 54 percent for the textbook and for voluntary reading 62 percent in grade 3. In grade 12 the optimal cloze scores were found to be 48 percent for textbooks and 36 percent for voluntary reading (Bormuth 1971, 138-139).

What to do if a textbook has been found to be too difficult for students? There are two possibilities: to rewrite the text in a more readable manner and/or reduce the amount of study material. The rules for readable writing have been presented by many authors (Baumann, Geiling, Nestler 1987; Flesch 1960; Klare 1985; Mikk 1984) and we will not refer to the rules here. We will illustrate the calculations of the optimal amount of study material using two examples. The idea laid down as the foundation for the calculations is that the amount of compulsory study material should be reduced to the extent that enables to achieve a positive mark by almost all the students.

In the first example the results of a test in geography are used as the indicator of textbook difficulty. The test was written by 854 ninth grade pupils who used the textbook (V ja IX... 1974)*. The test results were assessed on a 20-point scale (Table 1)

Table 1

The results of the test in geography

Score	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Number of testes	1	3	8	6	22	24	31	44	59	73	80	67	103	80	67	59	53	30	29	15
Percentage of testes			1	1	3	3	4	5	7	9	9	8	12	9	8	7	6	3	3	2
Cumulative percentage			100	99	98	95	92	88	83	76	67	58	50	38	29	21	14	8	5	2

The results reveal that the pupils, to be more exact – 95 percent** of them, knew the material well enough to score six points. That means that the pupils should be given a satisfactory mark for a six-point score and, consequently, six points should represent the knowledge of half of the appropriate material. (According to the grade programme a satisfactory mark presupposes that at least half of the material has been learnt). The full amount of appropriate material in this text will be equivalent to 12 points. In other words, judging by the results of this test, the degree of efforts required by the geography programme and the textbook is to be cut by (20–12)/20 · 100 percent = 40 percent.

As we see from the example, the results of a representative test can be easily used to calculate the amount of the study material appropriate for the representative group of students. The calculation can be made more precise if the model of frequency distribution of the test results, especially the end of the smaller values, is used. Let us have another example.

Students of eighth grade scored on average 26.1 points in a test on anatomy in the 1978/79 school year. The standard deviation of their results was 8.1 points and the possible maximum number of points was 42. The results are depicted in Figure 1.

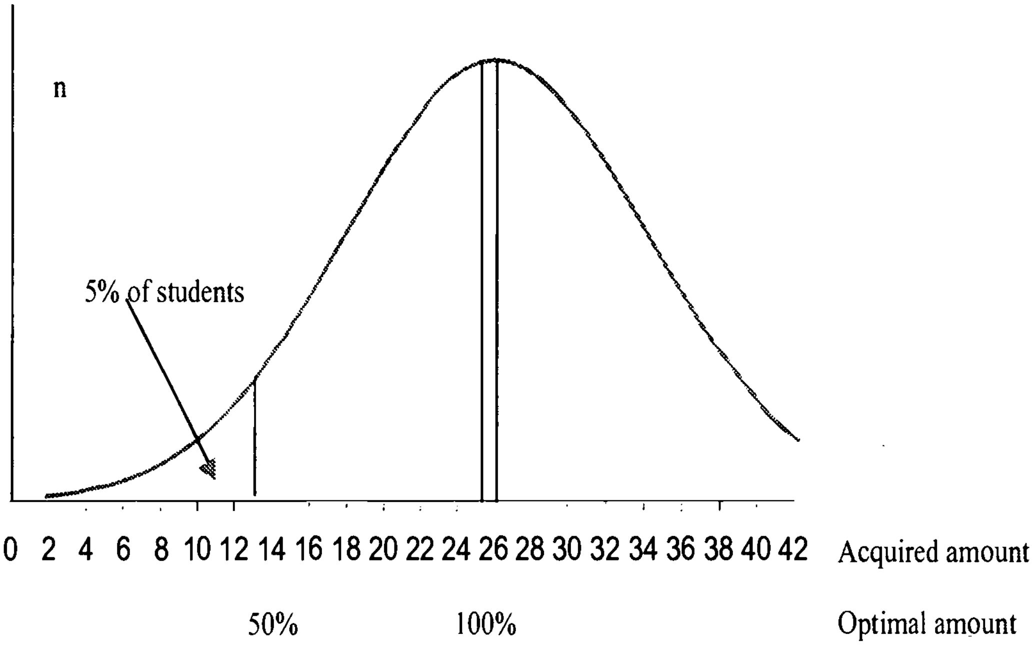

Figure 1. Calculating the optimal amount of the study material relying on the results of a textbook valid test carried out in a representative sample of students.

As in the previous example, I suppose that in our calculations 5 percent of students can have an unsatisfactory mark. I also suppose that the distribution of students is normal. According to the characteristics of normal distribution, 5 percent of the results are always lower than an average result minus 1.64 standard deviations. In our example, 5 percent of the students had a result lower than 26.1 – 1.64 · 8.1 = 12.8 points. The last value should be a boundary line between unsatisfactory and satisfactory marks or, in other words, it denotes a half of the optimal amount of the study material. The optimal amount corresponds to 2 · 12.8 or 25.6 points. Consequently, the amount of study material in the anatomy textbook should be reduced to (25.6/42) · 100 percent = 61 percent of its original amount.

The calculations above can be written in a generalised form as follows

$$OA = \frac{2(\overline{X} - 1.64\delta)}{X_{max}} 100\% \tag{7.4}$$

where

OA – the optimal amount of the study material expressed in the percentage from the real amount,

X – the mean result of the representative sample of students in the textbook valid test,

* – the standard deviation of the results,

Xmax – the possible maximum result in the test.

The Formula (1) is open for discussion in many aspects but the idea of calculating the optimal amount of the study material relying on the results of the students' learning is sound. The teachers use the idea in their everyday work: they reduce the amount of the study material if their students cannot acquire it appropriately, and the teachers accelerate learning if their students acquire the material on very high levels. Textbook authors cannot rely on the results of individual students. They must use the results of a representative sample of students.

Experimental comparison of the quality of two textbooks

The experimental comparison of two textbooks is based on some indices of the efficiency of the textbooks but the values of the indices depend on a broad variety of factors. Here is a list of some of them.

1. Students: socio-economic status, abilities, motivation, prior-knowledge of the topic, diligence, health, etc.
2. Teachers: professional competence, attitudes towards teaching, diligence, etc.
3. Textbooks: content, comprehensibility, illustrations, learning methods, etc.
4. Tests to measure effects: difficulty of questions, time to answer, etc.

If there are so many factors of the efficiency of learning, how can we decide which part of the results is due to the textbooks and which part is caused by other factors?

The problem can be solved by equalising the conditions of using the two textbooks for comparison. If the conditions are equal, then all the differences in learning results are due to the different quality of the textbooks. However, the simple idea is difficult to put into practice. We will look at it in some details.

There are some possibilities to *equalise the students* working with the two textbooks.

1. Students in both groups should be representative to the whole population of students. This is the most exact and the most expensive way to equalise the groups working with the textbooks for comparison. We have discussed the representativeness above.
2. The same students work with the two textbooks. It is possible if the content of the textbooks is different but usually two textbooks of the same content are compared and therefore the possibility is seldom used.

3. Students' prior knowledge, abilities, and other characteristics are measured and the results of studying the textbooks of those students are considered who match to the students in another group. It means, for example, that the results of some most capable students in the more capable group will not be considered while there are not so many capable students in another group. In this case, the comparable groups of students working with different textbooks will have the same average level of abilities and the same distribution of abilities.
4. An experiment can be carried out in the form of crossing groups (Latin Square). We will discuss it later.

Teachers' characteristics are even more difficult to equalise than the students' ones. In principle, the above mentioned approaches can be used but they are very difficult to realise in practice. For example, composing of two representative samples of teachers for the experiment is almost impossible. Certainly, teachers' professional competence, attitudes, etc. can also be measured and the teachers in two groups matched but I have never read about such practice. The experiment of crossing groups seems to be the simplest possibility to equalise teachers' characteristics working with the two textbooks.

One or more *textbook's characteristics* constitute the independent variable and textbooks must differ in this aspect. Textbooks may differ in many aspects and then the question arises which of them is crucial in determining different results of learning. To answer the question precisely, the characteristics not under study should be equal in comparable textbooks or their influence should be considered by covariance analysis.

The tests put to students after studying the textbooks should be representative to the textbooks. If the textbooks have the same study aims, then the tests will be the same for both of the textbooks. We discussed the composing of a representative sample of test items in the previous section.

After the short overview of the possibilities to equalise the conditions of working with the two textbooks let us look at *the experiment of crossing groups* in detail. This experiment is carried out in two parts (Table 2).

In the first part of the experiment textbook I is used by group A and textbook II is used by group B. In the second part of the experiment group A learns using textbook II and group B - textbook I. After completing both parts of the experiment the results of the learning process are assessed. In the whole experiment, textbook I is used by all the students and textbook II is used by the same students. Consequently students' and teachers' factors of learning efficiency are even for both textbooks. Differences in learning outcomes are due to differences in the textbooks.

Table 2

Outline of an experiment of crossing groups

Part of experiment	Group of Students	
	A	B
First	Textbook I	Textbook II
Second	Textbook II	Textbook I

There is some possibility that teachers are more enthusiastic in teaching with the new textbook than with the traditional one. To eliminate the influence, the experiment of crossing groups can be sometimes carried out in one room. Half of the students in the room use a new textbook and the other half uses the traditional one. Teacher's explanations are the same for all the students. In the second part of the experiment students exchange their textbooks.

The experiment of crossing groups is a good method for equalising the teachers' and students' factors in studying textbook efficiency. The method guarantees reliable results even if participants in the experiment are not strictly representative of the whole sample. At the same time, the experiment of crossing groups has a shortcoming. If, in the first part of the experiment, a study skill is acquired from one textbook, then the study skill enhances the results of learning in the second part of the experiment by the students who are using the second textbook. Some positive effect of the first textbook is misleadingly ascribed to the second textbook due to the experiment design. The experiment of crossing groups is not applicable if study skills, motivation or other effects that influence learning outcomes in the second part of experiment are considered. The experiment of crossing groups is usable when the acquisition of knowledge is the main aim of learning.

Let us have an example of using the experiment of crossing groups. The aim of the experiment was to evaluate the effect of using suggestions for understandable writing in Estonian. The investigation was carried out on textbooks of physics for grade 7 (14-year-old students) and anatomy for grade 8. Two chapters from both textbooks (about 80 pages) were rewritten according to the suggestions for understandable writing and the rewritten parts of textbooks printed as booklets. 2167 students participated in the experiment. After the first part of the experiment was over the results of learning were measured and the textbooks exchanged. The indices of the prior knowledge, text comprehension, its acquiring, information gain, and the persistence of knowledge were used to characterise the efficiency of learning. The measurements were repeated after the second part of the experiment. Summary results of the experiment are given in Table 3.

Table 3

Efficiency of suggestions for understandable writing

Indicator of efficiency	Level of the indicator (in percentage)		Efficiency percentage*
	Traditional textbook	Revised textbook	
Comprehension	64.6	73.1	13
Acquisition	55.5	63.0	13
Information gain**	40.4	45.9	14
Persistence of knowledge***	37.6	40.0	6

* The percentage is calculated consodering the levelresults in the group working with traditional textbook for 100 %.

** The maximum possible information gain is equaled to the acquiring of the text minus prior knowledge.

***The persistence of knowledge is calculated considering the level of acquiring for 100 %.

In the table, we see that all the indicators of learning efficiency ere higher when using revised texts. Following the suggestions for understandable writing enhanced learning efficiency by about 13%. All the effects were statistically significant.

Unexpected results of experimental research

Experimental research is aimed at verifying an hypothesis, for example, the new textbook is better than the previous one. Different data are collected and analysed to prove the hypothesis. The analysis is directed by a single goal - the hypothesis. At the same time, the data reflect the richness of the real world, and therefore they depict the other regularities as well. It is extremely useful to look at the data from some other points of view – some unexpected discoveries may be made.

The idea is known as secondary data analysis (Reeve & Walberg 1997) – the data gathered by one researcher for his/her purpose are reanalysed by another researcher to solve his/her problem. To enable the reanalysis, the data should be very well documented and may contain some aspects that are not needed for the first research. The data may function as data banks in sociological research (Anderson & Rosier 1997) accessible for other researchers as well. The data collector may look at them from many points of view and this fosters gathering more information than needed to answer the initial research question.

The first example of unexpected results is related to readability formula development. Our (the resurs was carried out together with Jaanus Elts and Toomas Tamman) aim was to develop a readability formula for biology texts in Russian. We took 48 texts from popular-scientific books on biology. The texts were about one typewritten page long. The texts were studied by 124 pupils of the 7th, 8th and 10th forms in Russian speaking schools in Estonia. All the pupils were asked to answer questions on the content of the text (to measure their level of prior knowledge), to read the texts, fill in a questionnaire, and answer another set of questions on the content of the text. The questionnaire included questions if the text was interesting for them (2) or not (1).

All the texts were computer-analysed. The analysis included the following aspects. 1. Establishing the distribution of words by their length, the distributions of sentences by their length and by other simple characteristics. 2. The morphological analysis of the words of the texts using programs which had been worked out by N. A. Dartschuk and her colleagues in Kiev (Automatisation ..., 1984). The morphological analysis determined the principal form of every word in the text, the part of speech to which the words in the text belonged and their frequency of occurrence. 3. The frequency rank of the words in our texts was established by comparing them with the entries of the frequency dictionary of Russian which we had been given by D. Buchstab from Moscow University. 4. As the abstractness of nouns and the number of terms in texts greatly affects text comprehension, the degree of abstractness of every noun in the text and their role as terms in the text were assessed by human experts (Elts 1992).

The arithmetical mean values of the pupils' answers were correlated with the characteristics of the texts. Some of the correlation coefficients have been presented in Table 4.

The aim of our research was to find the correlation coefficients given in the last column of Table 4. The correlation coefficients met our expectations. The computations on computer are easy and we calculated the correlation coefficients in the last but one column as well. The coefficients bewildered us! For example, the first of them (-0.53) means that before reading the text, the students answered fewer questions correctly on the texts that had longer sentences. How can it be? How did the students know in which texts the sentences were longer and answered fewer questions correctly before reading the texts? It can not be! Our experiment has fully mysterious results! No scientific conclusion can be drawn!

Table 4

Validity of some text characteristics in predicting reading outcomes

No	Characteristic	Average	Standard deviation	Correlation coeficient* with		
				Interest in reading No. 202	Pre-test score No. 210	Post-test score No. 212
1	2	3	4	5	6	7
26.	Proportion of sentences of 40 or more letter spaces**	0.92	0.10	-0.55	-0.38	-0.63
29.	Proportion of sentences of 70 or more letter spaces	0.76	0.20	-0.68	-0.48	-0.75
30.	Proportion of sentences of 80 or more letter spaces	0.69	0.22	-0.71	-0.51	-0.74
31.	Proportion of sentences of 90 or more letter spaces	0.61	0.24	-0.70	-0.53	-0.73
35.	Proportion of sentences of 130 or more letter spaces	0.36	0.23	-0.62	-0.34	-0.59
78.	Proportion of words of 9 or more letters	0.26	0.07	-0.76	-0.53	-0.78
90.	Number of letter spaces in sentence	119	36	-0.66	-0.44	-0.65
91.	Number of letters in word	6.3	0.6	-0.75	-0.54	-0.76
97.	Frequency of the text's words in the SLD***	968	205	0.55	0.37	0.56
103.	Frequency of the text's nouns in the SLD	26.6	17.2	0.50	0.58	0.48
104.	Repeating rate of the nouns in the text	1.35	0.13	-0.48	-0.42	-0.47
109.	Percentage of nouns in the text	34.8	5.0	-0.63	-0.49	-0.62
233.	Mean terminological index of nouns	1.53	0.24	-0.64	-0.45	-0.56
236.	Percentage of abstract nouns	23.6	16.0	-0.70	-0.49	-0.71
242.	Percentage of terms which are not used in everyday speech	13.4	11.2	-0.71	-0.54	-0.64

* Correlation coefficients with the absolute value o.29 or above are statisically significant at 0.95 level.

** Letter spaces are all letters, punctuation marks, and gaps betwen words.

*** SLD – spoken language dictionary composed in Moscow University by Buchstab and colleagues.

It took a year to understand the correlation coefficients. The correlation coefficients are based on the fact that some topics are better known in society than others. If a topic is known, then people write about it in shorter sentences, using more familiar and less abstract words. Also students know the topic better and give more correct answers before reading the text that the author has written in a simpler way. Both of the correlated characteristics - the level of correct answers before reading the text and the readability level of the text, have one predictor variable – the familiarity level of the topic in society. Or in other words, readability formulae measure the level of familiarity of the text content to the readers to a certain degree.

Afterwards we divided the biology texts into three categories: texts about microbiology, organisms and ecology. The texts on organisms had, on average, the shortest sentences, the shortest words, the more words frequent in everyday speech, and the least abstract words. The same texts were evaluated as most interesting among the three categories of the texts and the students could give the best answers to the questions about their content before reading the texts (Elts & Mikk 1993).

Analysing the data of the above described experiment, we paid attention to the fact that the coefficient of correlation (in the last column of table 4) changed systematically if the dividing line between the short and the long sentences changed. The correlation coefficient between the results of the post-test score and the percentage of long sentences was low (in absolute value) when the relatively low value of sentence length was taken as the dividing line between the short and the long sentences. If we moved the dividing line from the short sentences to the longer sentences, then the correlation coefficient first rose to the maximum (-0.75) and then began to decrease. What might be the reason for these systematic changes? Answering the question, we found a new and exact method to elaborate the optimal value of sentence length for different students (Elts & Mikk 1996). These examples prove that unexpected results of experiments may be even more important than the initial aim for collecting data.

Conclusion

What really matters is the efficiency of using textbooks at school. The results can be most validly measured by experiments in schools. Therefore the experimental method of measuring textbook quality deserves the special attention of textbook researches. The experimental method is even more important because it enables to validate the different ways of textbook analysis, and experts also rely on the results of field testing in formulating their opinions about textbooks.

The experimental method is the most complicated method in textbook research. The researcher has to think about the representativity of students' groups, or about their equality, about the validity of experimental design and measurements etc. The ideal conditions can hardly be achieved for different reasons, or as M. J. Lawson (1977, 133) put it: "Experimental design requires juggling of ideals and practicalities". The more deviations from the ideal, the less valuable are the results.

An experiment is the most expensive method of textbook investigation. It should be used after the textbook analysis and the removal of the shortcomings which were elicited in the analysis. Otherwise in the experiment we give the students a textbook with shortcomings that hinders their development and this contradicts the ethical norms of educational research.

Usually the measurement of textbook efficiency is many-sided in experiments: content tests, questionnaires to teachers and students, the survey of students work, etc. The collected exact and thorough data enable the researcher to solve different problems and sometimes serve as a basis for unexpected discoveries.

Bibliography

Agur U. Toim K., Unt I. (1967). *Programmõpe ja õpimasinad [Programmed learning and teaching machines].* Tallinn: Valgus, 320 lk.

Anderson J., Rosier M. J. Data banks and data archives. In: *Educational Research, Methodology, and Measurement: An International Handbook.* Second edition. Ed. By John P. Keeves. Oxford, New York, Tokyo: Elsevier Science 1997, pp. 344-349.

Anderson L. W., Block J. H. (1985). Mastery learning model of teaching and learning. In T. Husen, T. N. Postlethwaite. (eds.) *The International Encyclopaedia of Education:* Research Studies. Oxford, 3219–3220.

Avtomatizatsiya analiza nauchnogo teksta (1984). *[Automation of science text analysis].* In Verbickii, Darchuk et al. Kiev: Naukova Dumka, 258 p. (in Russian).

Babanskii Y. K. (1977). Optimizatsiya protsessa obucheniya. Obshchedidakticheskii aspekty [*Optimisation of learning process. General didactic aspect].* Moscow: Pedagogika, 256 p. (in Russian).

Baumann M., Geiling U., Nestler K. (1987). Katalog verständnishemmenden Textmerkmale ("Störstellenkatalog"). *Informationen zu Schulbuchfragen.* Berlin: Volk und Wissen Volkseigener Verlag, Heft 56, 36-55.

Bormuth J. R. (1968). The cloze readability procedure. *Elementary English,* vol. 45, 429-436.

Bormuth J. R. (1971). *Development of standards of readability: Toward a rational criterion of passage performance.* Chicago: The University of Chicago, U.S. Department of Health, Education and Welfare. Manuscript, 151 p.

Chall J. S. (1958). *Readability. An appraisal of research and application.* Columbus, Ohio: Ohio State University Press, 202 p.

Cherepanov V. S. (1991). Teoreticheskie osnovy pedagogicheskoi ekspertizy [Theoretical basis of educational expertise.] Moscow: NII TIP. (Unpublished doctoral dissertation) (in Russian).

Cristoffersson N. O. (1971). *The Economics of Time in learning.* Malmö, 185 p.

Die Schulbuchbegutachtung in verschiedenen Ländern. (1991). Wien: Institut für Schulbuchforschung, 30 S. (Manuskript).

Elts J. (1992). A readability formula for texts on biology. Psychological problems of reading. Theses of papers for the international scientific conference. Vilnius, 42–44.

Elts J., Mikk J. (1993). Kompliziertheit der Texte zur Umwelterziehung. In: Aus dem Wissenschaftlichen Leben der Pädagogischen Hochschule Halle-Köthen. Heft 2: Schulbuch und Umwelterziehung, S. 55-62.

Elts J., Mikk J. (1996). Determination of optimal values of text characteristics. *Journal of Quantitative Linguistics,* vol. 3, no. 2, 144–151.

Entin E. B., Klare G. R. (1985). Relationship of measures of interest, prior knowledge and readability to comprehension of exposition passages. Advances in Reading/ Language Research, vol. 3, 9–38.

Flesch, R. (1960). *How to write, speak and think more effectively.* New York: Harper & Brothers, 362 p.

Green K. (1984). Effects of item characteristics on multiple-choice item difficulty. *Educational and Psychological Measurement,* vol. 44, 551-561.

Hater M. A., Kane R. B. (1970). The cloze procedure as a measure of the reading comprehensibility and difficulty of mathematical English. Purdue University, Lafaette, Ind. 24 p.

Klare G. R. (1985). *How to write readable English.* 5th ed. London, etc. Hutchinson, 144 p.

Kuznetsov O. A., Khromov L. M. (1977). Tekhnika bystrogo chteniya [Technique of speedy reading]. Moscow: Kniga, 192 p. (in Russian).

Lawson M. J. Experimental studies. In: *Educational Research, Methodology, and Measurement: An International Handbook.* Second edition. Ed. By John P. Keeves. Oxford, New York, Tokyo: Elsevier Science 1997, pp.126-134.

Maanso V. (1969). Vaagimist vajavad nõuded lugemisoskusele keskastmes [Requirements to reading ability in middle grades need discussion]. Nõukogude Kool, no. 5, 337-344 (in Estonian).

Mikk J. (1981). Teoriya izmereniya i optimizatsii stepeni slozhnosti uchebnogo materiala v obshcheobrazovatelnoi shkole [Theory of the measurement and optimisation of the degree of complicacy of the study material in comprehensive school]. Doctoral dissertation. Manuscript. Tartu, 434 p. (in Russian).

Mikk J. (1984). Empfehlungen für die Verbesserung der Verständlichkeit des Lehrtextes. *Informationen zu Schulbuchfragen.* Heft 48, S. 97-121.

Nikandrov N. D. (1970). Psikhologo-pedagogicheskie voprosy metodiki sostavleniya programmirovannykh materialov v rabotakh zarubezhnykh programmistov [Psychological and pedagogical problems of the methods for compiling programmed teaching materials in the publications of foreign programmists]. Moscow: Znanie, 46 p. (in Russian).

Potter T. C. (1968). A taxonomy of cloze research. Part I. Readability and Reading Comprehension. Manuscript. 50 p.

Rapoport I. A., Gohlerner M. M., Selg R., Sotter I. (1976). O diagnosticheskikh funktsiakh testovoi metodiki dopolneniya [Diagnostic functions of cloze procedure]. *Inostrannye yazyki v shkole,* no. 2, 31-37 (in Russian).

Rauch M., Tomaschewki L. (1986). Reutlinger Raster zur Analyse und Bewertung von Schulbücher und Begleitmedien. Reutlingen, 64 S.

Reeve R. A., Walberg H. J. Secondary data analysis. In: Educational Research, Methodology, and Measurement: An International Handbook. Second edition. Ed. By John P. Keeves. Oxford, New York, Tokyo: Elsevier Science 1997, pp. 439-445.

Sepetliev D. (1968). *Statisticheskie metody v nauchnykh meditsinskikh issledovaniyakh* [Statistical methods in scientific medical research]. Moskow: Medicina, 419 p.

Smith H. P., Dechant E. V. (1961). *Psychology in teaching reading. 2nd print.* New York, Englewood Cliffs, 470 p.

Taking the SAT I. Reasoning test. (1994). The College Board, Educational Service, 80 p.

Talyzina N. F. (1975). *Upravlenie protsessom usvoeniya znanii (psychological foundations) [Leading the process of acquiring knowledge (psychological foundations)].* Moscow: Moscow University Press, 343 p. (in Russian).

Taranov L. N. (1976). *Optimizatsiya ponimaniya uchebnogo materiala v usloviyakh programmirovannogo obucheniya [Optimisation of the comprehension of educational material in programmed learning].* Kiev, 22 p. (in Russian).

Tholey M. (1996). *Bezugsrahmen Deutsch. Dimensionen, Kategorien, Items, Spezifizierungen.* Enschede. 14 S.

Tyson-Bernstein H. (1989). Textbook development in the United States: How good ideas become bad textbooks. In J. P. Farrell & S. P. Heyneman. (eds.) *Textbooks in the Developing World. Economic and Educational Choices.* Washington: The World Bank, 72-87.

V ja IX klassi geograafia 1973. a. kontrolltööde tulemused. (Results of a test on geography in fifth and ninth form in 1973.) Tallinn: ENSV Haridusministeerium, 50 p. (in Estonian).

Vassilchenko L. (1995). Students' selfrating: possibilities of its application in the study of information conditions of the learning process. *Family and Textbooks.* Tartu: University of Tartu, 61-77.

Weintraub S. (1968). The cloze procedure. *The Reading Teacher,* vol. 21, no. 6, 567, 569, 571, 607 pp. March.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Experimental evaluation of textbooks and multimedia	
Author(s): Jaan Mikk	
Corporate Source: Staffan Selander & Marita Tholey (Eds.). New educational media and textbooks. Stockholm Institute of Education Press	Publication Date: 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

Level 1	**Level 2A**	**Level 2B**
+		
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

Signature: J. Mikk	Printed Name/Position/Title: Jaan Mikk. Professor of Education. Doctor of Education	
Organization/Address: Department of Education University of Tartu, Ülikooli 18, 50090 Tartu, Estonia	Telephone: +07 375156	Fax: +07 375 156
	E-mail Address: jmikk@ut.ee	Date: 17 May 2002

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address: