The book cover features a stylized, high-contrast illustration. In the foreground, two faces are depicted in profile, facing each other. The face on the left is a woman with dark, wavy hair, and the face on the right is a man with dark hair. They are rendered in a graphic, almost woodcut style with bold lines and flat colors. Behind them, a large, stylized figure of a person is visible, composed of many small, repeating human silhouettes. To the left of the faces, a large, stylized letter 'T' is prominent, with the word 'TEST' written across it in a bold, sans-serif font. The background is a deep blue with a pattern of diagonal lines. The entire cover is framed by a decorative border of small, repeating triangles.

UNDERSTANDING EDUCATIONAL MEASUREMENT

ERNEST MCDANIEL

UNDERSTANDING EDUCATIONAL MEASUREMENT

ERNEST MCDANIEL
Purdue University



Boston, Massachusetts Burr Ridge, Illinois Dubuque, Iowa
Madison, Wisconsin New York, New York San Francisco, California St. Louis, Missouri

McGraw-Hill

A Division of The McGraw-Hill Companies

Book Team

Managing Editor *Sue Pulvernacher-Alt*
Production Editor *Karen A. Pluemer*
Visuals/Design Developmental Consultant *Marilyn A. Phelps*
Visuals/Design Freelance Specialist *Mary L. Christianson*
Marketing Manager *Elizabeth Haegele*
Advertising Manager *Nancy Milling*

Executive Vice President/General Manager *Thomas E. Doran*
Vice President/Editor in Chief *Edgar J. Laube*
Vice President/Sales and Marketing *Eric Ziegler*
Director of Production *Tickie Putman Caughron*
Director of Custom and Electronic Publishing *Chris Rogers*

President and Chief Executive Officer *G. Franklin Lewis*
Corporate Senior Vice President and Chief Financial Officer *Robert Chesterman*
Corporate Senior Vice President and President of Manufacturing *Roger Meyer*

Cover Illustration by Wordsworth Illustration.

Cover and interior designs by Terri W. Ellerbach.

Illustrations by Wordsworth Illustration unless noted otherwise.

Copyedited by Karen Dorman

Copyright © 1994 by Wm. C. Brown Communications, Inc. All rights reserved

Library of Congress Catalog Card Number: 93-70908

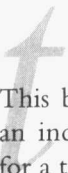
ISBN 0-697-13208-0

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

Printed in the United States of America

10 9 8 7 6 5 4 3 2

Preface



This book stems from the belief that there is an increasing need within teacher education for a text that introduces students to the essentials of measurement and evaluation using a style that injects human interest, imagination, and humor into the subject matter. This book will focus exclusively on topics essential to effective classroom teaching and will treat these topics in a manner that emphasizes the human context that surrounds and guides their use. In addition, this book is intended to help students overcome their perception of measurement as an abstract activity interesting only to those who are mathematically inclined.

A number of admirable texts are available that treat measurement topics thoroughly and competently. I have cited many of these texts throughout the chapters of this book, and the reader will profit from exploring these sources more fully. This book, rather than attempting to be comprehensive and detailed, drives home a few essential concepts and skills through a focused presentation and a style that I hope will be seen as friendly and engaging.

There are, to be sure, some statistical formulas, but the emphasis is always on the conceptual material, not the arithmetic operations. Students who enjoy number

crunching can apply the formulas to specific examples, and those who want to skip these passages may do so without feelings of guilt or concern that they are missing building blocks in their conceptual development. Dozens of students who have used this text in draft form have confessed to “math fright” at the beginning of the course and returned later to express gratitude for the good foundation they acquired for later courses in statistics.

Educational measurement is an intensely human activity. No test ever arrives in the classroom independent of numerous decisions made by people on the basis of what they know, believe, and value. I like the human context in which measurement decisions are made. I see teachers as valuing broad goals in education and welcoming ways of designing tests to incorporate a philosophy of education with which they feel comfortable. Thus, readers will find an undercurrent of applause running throughout this book for teachers moving away from pure textbook learning and toward the development of thinking processes and positive attitudes toward self and subject matter. From a measurement perspective, this concern is expressed in an emphasis on higher cognitive processes in achievement tests and in

a direct treatment of the measurement of thinking growing out of my own research in this area. Newer trends in performance assessment and portfolio assessment also represent, at base, an interest in humanizing education in the sense of becoming more responsive to individual ways of growing. These approaches to evaluation encourage students to become more reflective and self-actualized.

There is a second aspect of testing as a human endeavor that is totally neglected in most measurement texts. As authors of texts, we are too busy teaching how the IQ score is computed to provide even a glimpse of the men and women who struggled to solve measurement problems. For example, there is Alfred Binet, the French psychologist with a passionate interest in understanding everything about the human mind, a fascinating personality who spent long hours giving tests to his two daughters, invented a mechanical device for recording the artistry of piano players, wrote really macabre plays for the Paris theater, and attacked many of the myths that were currently held about indices of character and personality. His colleagues were chagrined when he demonstrated that they could not, on the basis of handwriting, tell the difference between prominent Frenchmen and criminals.

I have included an entire chapter on Binet in this book, and where appropriate, have provided brief biographical comments on some of the other figures in testing. It seems amazing that in other academic fields we recognize a Pasteur, a Mendel, a Mach, a Faraday; but in testing, we somehow see tests of intelligence, aptitudes, or interests appearing on the scene

as disembodied entities, separated from their authors' personal motivations and theories of human abilities.

The story of testing is a human story reaching back to Sir Francis Galton, and we will see in a moment that you are linked to Galton through a handshake that spans a hundred years. Galton established a laboratory in Keniston, England, in 1884, where he measured people for a variety of human traits in the interest of answering questions about human abilities. An American, James McKeen Cattell, visited Galton in his laboratory and we can imagine at the parting, Cattell extended his hand and said, "It has been a pleasure." Cattell returned home to conduct a series of researches in testing, and was the first American to use the term "mental test" in a journal article. In due time, Cattell recruited a young man who had been his son's roommate in college to come to New York and assist him in compiling data about prominent American men of science. Fredrick Kuder performed this work and went on to make his own significant contributions to testing. As Kuder left Cattell, we can imagine that he extended his hand and said, "It's been a pleasure." At the annual meeting of the American Psychological Association in Los Angeles in 1984, Kuder reminisced about his professional experiences and after the talk, Ernest McDaniel, a professor of measurement at Purdue University, extended his hand and said, "It has been a pleasure." And now, to all readers, I extend my hand to welcome you to the field of educational measurement. I hope you will like it.

Acknowledgments

I am deeply indebted to the following individuals who reviewed the material for this text and offered suggestions that resulted in a number of significant changes in the final book:

Ralph F. Darr, Jr. University of Akron
Patricia A. Cook University of
Indianapolis
Robert E. Bonner Bartlesville Wesleyan
College
Landa Trentham Auburn University
Glen Nicholson University of Arizona
T. Patrick Mullen California State—
San Bernardino
Ernest Davenport University of
Minnesota

I wish to thank Mark McDaniel, who teaches cognitive psychology at Purdue, for his suggestions on the chapter on measuring intelligence.

Paul Tavenner of Brown & Benchmark Publishers saw the manuscript through its developmental phases, and I appreciate his efforts and those of the many specialists who contributed their skills to the final publication.

Jill Brady competently handled a variety of secretarial tasks and word processing activities associated with the development of the manuscript.

I am also grateful to the many students who identified errors and made suggestions. Their responsive interaction with the material reinforced my belief that the subject of testing can be interesting and enjoyable.

Brief Contents

part one

Principles of Measurement

chapter one

An Introduction to Educational Measurements 1

chapter two

Elementary Statistical Concepts 17

chapter three

Reliability 43

chapter four

Validity 67

chapter five

Norms, Standardization Procedures, and Expectancy Tables 83

chapter six

Selecting and Critiquing Standardized Tests 111

part two

Measuring Educational Achievement

chapter seven

Constructing Classroom Tests 129

chapter eight

Item Analysis and Item Revision 151

chapter nine

Interpretive Exercises and Essay Tests 165

chapter ten

Performance Assessment 181

chapter eleven

Measuring Thinking Processes 205

part three

Measuring Human Behavior

chapter twelve

Binet and His Great Invention 219

chapter thirteen

Measuring Intelligence 233

chapter fourteen

Measuring Aptitude 265

chapter fifteen

Measuring Interests 281

chapter sixteen

Measuring Affective Aspects of Schooling 295

chapter seventeen

New Trends in Testing 321

Contents

Preface xiii

Acknowledgments xv

chapter one

An Introduction to Educational Measurements 1

Principles of Measurement 2

Designing and Constructing Classroom Tests 4

Measuring Human Behavior 6

New Trends in Testing 8

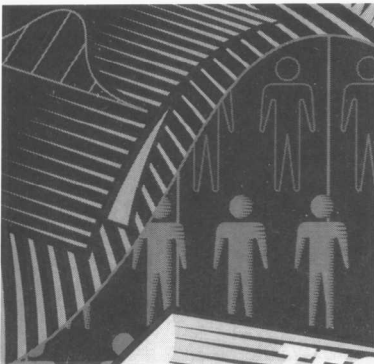
Ending Comments 10

Summary 12

References 13

part one

Principles of Measurement



chapter two

Elementary Statistical Concepts 17

Snow and Statistics 18

Types of Scales 19

Frequency Distributions 21

Measures of Central Tendency 24

Measures of Variability 27

The Normal Curve 32

Going From Standard Deviations to Standard Scores 35

The Correlation Coefficient 37

A Word About Inferential Statistics 40

Summary 42

References 42

chapter three

Reliability 43

The True Score Plus Error 45

Sources of Measurement Error 45

Procedures for Determining Test Reliability 46

Reliability Coefficients Influenced by Group

Characteristics 55

How High Should a Reliability Coefficient Be? 55

Using Reliability Coefficients 56

The Standard Error of Measurement 58

Correction for Attenuation 63

Summary 64

References 65

chapter four

Validity 67

- Approaches to Test Validity 69
- How High Should a Validity Coefficient Be? 74
- Reading the Manual 76
- Validity Is Related to Purpose 79
- Summary 80
- References 81

chapter five

Norms, Standardization Procedures, and Expectancy Tables 83

- Raw Scores 84
- Criterion-Referenced Tests 85
- Norm-Referenced Tests 86
- An Exercise in Using Norms 87
- Selecting the Group for Norming the Test 91
- The Norms 94
- Derived Scores 94
- Expectancy Tables—Better Than Norms 104
- Norms Are Not Normal 107
- Summary 109
- References 110

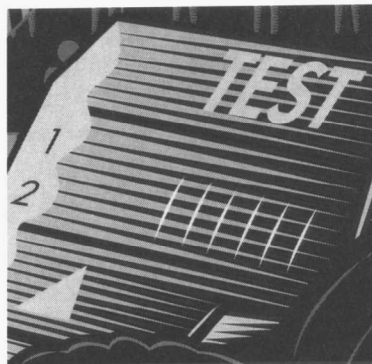
chapter six

Selecting and Critiquing Standardized Tests 111

- The Mental Measurements Yearbooks 112
- Evaluating Tests from Data in the Test Manual 115
- Evaluation of the Stanford Early School Achievement Test, Third Edition 120
- Summary 124
- References 125

part two

Measuring Educational Achievement



chapter seven

Constructing Classroom Tests 129

- Tests Reflect Teachers' Objectives 130
- The Bloom *Taxonomy of Educational Objectives* 131
- The Table of Specifications 133
- Writing Test Items 138
- Objective Test Items 144
- Summary 149
- References 149

chapter eight

Item Analysis and Item Revision 151

- The Rationale of Item Analysis 152
- Item Analysis of Classroom Tests 154
- How Difficult? 157
- How Discriminating? 157
- Item Revision 158
- Machine Generated Item Statistics 159
- Building an Item File 161
- Item Analysis of Criterion-Referenced Tests 161
- Item Response Theory 162
- Summary 164
- References 164

chapter nine

Interpretive Exercises and Essay Tests 165

- Expressive Objectives and Convergent Thinking 166
- The Interpretive Exercise 167
- Essay Tests Versus Multiple-Choice Tests 169
- A Bit of History 172
- Unreliability of Essay Tests 173
- Problem of Sampling and Setting Adequate Questions 175
- Advantages of Essay Tests 175
- Writing Essay Questions 176
- Scoring Essay Tests 177
- Summary 179
- References 180

chapter ten

Performance Assessment 181

- Reasons Behind Performance Assessments 183
- Performance Tasks as Realistic Work Samples 184
- Scoring of Performance Tasks 186
- Psychometric Properties of Performance Tasks 186
- Performance Assessment Applied by the Teacher 194
- Getting Started in Performance Assessment 198
- Using Simulations 199
- Summary 203
- References 203

chapter eleven

Measuring Thinking Processes 205

- Thinking in Education 206
- Standardized Tests of Critical Thinking 207
- Alternative Efforts to Measure Thinking Processes 209
- Levels of Cognitive Complexity—An Approach to the Measurement of Thinking 211
- Discussion of Levels of Cognitive Complexity 213
- Summary 214
- References 215

part three

Measuring Human Behavior



chapter twelve

Binet and His Great Invention 219

- The Early Scales 220
- The Stanford-Binet 225
- Looking Back 228
- Summary 230
- References 230

chapter thirteen

Measuring Intelligence 233

- Charles Spearman and The Theory of Two Factors 235
- Becoming More Thoughtful About Intelligence Testing 239
- Crystallized and Fluid Intelligence 240
- The Fourth Edition Stanford-Binet 241
- The Wechsler Intelligence Tests 246
- Wechsler Intelligence Scale for Children-Third Edition (WISC-III) 247
- Special Class Placement and IQ Scores 255
- Information Processing Approaches to The Measurement of Intelligence 256
- Summary 262
- References 263

chapter fourteen

Measuring Aptitude 265

- The Nature of Human Abilities 266
- Aptitude Tests 268
- The Differential Aptitude Tests 269
- The General Aptitude Test Battery (GATB) 274
- Armed Services Vocational Aptitude Battery (ASVAB) 276
- Creativity as a Specific Ability 277
- Summary 278
- References 279

chapter fifteen

Measuring Interests 281

- Edward Strong and His Vocational Interest Blank 282
- The Strong Interest Inventory 284
- The Kuder Preference Record 286
- The Self-Directed Search 289
- Measuring Interests at the Elementary Level 290
- Interpreting Interest Inventory Scores 292
- Summary 292
- References 293

chapter sixteen

Measuring Affective Aspects of Schooling 295

- Self-Report and Projective Techniques 297
- Using Existing Tests or Designing Your Own 298
- Finding Existing Instruments 298

- Constructing Measures of Perceptions and Attitudes 299
- Attitudes Toward Subjects and School 299
- Why Students Work: Motivational Orientation Scales 301
- Learning Styles and Learning Strategies 302
- Self-Concept Scales 305
- Measures of Individual Adjustment 307
- Nominating Techniques, Sociograms, and Direct Observations 309
- Summary 319
- References 319

chapter seventeen

New Trends in Testing 321

- Portfolio Assessment 322
- Dynamic Assessment 330
- Computerized Adaptive Testing 333
- Summary 336
- References 337

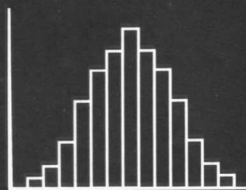
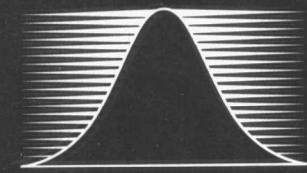
- Appendix* 339
- Glossary* 343
- Bibliography* 351
- Index* 359

CAMPUS REGISTRATION

Well, I have done it. . . . I've signed up for a course in educational measurements. I know it's going to be loaded with statistics and dry and boring, but I need it for my program. I wonder if I will get what I really want from this class. I would like to do a better job of interpreting scores of standardized tests. I would like to make better tests for my students and I would like to know what to think about such things as "intelligence" and "aptitude."

chapter one

An Introduction to Educational Measurements



Measurements and evaluations are interlinked with almost every activity of teaching. Teachers with a measurement background are able to construct, select, and use tests more skillfully than teachers without training in measurement. A text in measurement written for teachers should present the principles of measurement, which form the basis for selecting tests and interpreting results. It should explain how to write good classroom items and how to analyze the items after the test has been administered. It should include enough statistics so that test manuals and research reports can be read with understanding. It should also acquaint the reader with the major test instruments in the fields of academic abilities, interests, and personal adjustment.

With these objectives in view, this book is organized into three main sections: (1) understanding general measurement principles, (2) constructing and analyzing classroom tests, and (3) examining specific tests in the areas of intelligence, interests, aptitudes, and adjustment. This chapter introduces each of these three areas and provides an overview of the major ideas in each one.

Principles of Measurement

Measurements provide a precise rather than a vague basis for making educational decisions. A teacher, judging a student's mathematical ability as "pretty good," may assign extra work to do after regular homework is completed. A teacher, knowing that the student exceeds 97 percent of a national sample, may recommend a special program for mathematically gifted.

While measurement lends precision to the study of human behavior, tests may suffer a number of shortcomings. Tests seem scientific. Tests seem to emanate from highly authoritative sources. For these reasons, tests may be dangerous; they may lull us into the belief that we are using highly objective, scientific tools when we are not.

Teachers need to know what qualities to look for in evaluating tests and need to feel comfortable with the specialized terminology that professionals use when talking about tests.

Statistical Concepts

A brief introduction to statistical concepts provides the foundation for consideration of measurement principles. Notice we are not calling this section statistical calculations. Our concern is with the concepts. When specific calculations are needed, it has become easy to ask a computer to

perform these calculations. These exercises can be done with ease because the computer has already been loaded with statistical packages that automatically perform the needed calculations.

Descriptive statistics are used to describe a group of measurements such as a distribution of test scores. Some teachers visualize a distribution of scores as a landscape with each test score placed in its appropriate place, low scores to the left, high scores to the right. If you like, you can think of a bell-shaped, normal distribution curve as a mountain of answer sheets with a peak and two sloping sides.

Within this territory the middle point, usually the average or mean, is the key landmark. You can build a surveyor's tower there. From this surveyor's tower, you can see the extent of your domain, the range from end to end. You can also calibrate your surveyor's scope so when you spot a test score, you can tell how many "standard deviations" it is from your position at the mean. The standard deviation is the second landmark, and a solid understanding of it makes for easy transition into applied measurement terms.

Correlation coefficients are the indispensable tool of the measurement person. Correlations tell us the degree to which two variables are associated with one another. If we give a test twice to a group of students and correlate the scores, we have a reliability coefficient. If we correlate scores for math aptitude with math achievement, we have a validity coefficient. I realize these are terms that have not yet been defined, but I am illustrating how an understanding of correlations is a prerequisite to the following discussions about the qualities of a test. Let us now turn to these essential qualities of a good measurement device.

Validity, Reliability, and Norms

It is impossible to talk about tests or test scores without using a specialized vocabulary. Understanding this vocabulary involves more than definitions; it involves learning about a set of procedures that are employed to "test the test."

When evaluating a test, you will be seeking information relevant to three characteristics: reliability, validity, and norms. You may want to think of test evaluation as asking three major questions:

1. Reliability: Does the test give consistent scores on subsequent occasions?
2. Validity: Does the test measure what it claims to measure?
3. Standardization: Are the norms based on a representative sample?

These three criteria for judging tests will be appropriate whether the test was published in Iowa City or in Princeton. These criteria for judging tests will also be essential no matter what the test is designed to measure: reading readiness, for example, or self-concept, or teacher competency. Learning to use these three principles of measurement will enable you to be a better consumer of tests and a far better interpreter of test information.

The essential information for evaluating tests is in the test manual. Without some background in measurement, however, there may be a tendency to leaf through the manual vaguely hoping for some important information to leap off the pages. If you understand that the quality of a test depends on its reliability, validity, and the way it was standardized, then these headings in the manual become the signposts indicating that relevant information will follow. Although test terminology may appear technical, we will see that the terms describe common sense concerns raised by anyone asking the question: "How good is this test?"

Designing and Constructing Classroom Tests

Constructing classroom tests is one of the many applications of measurement principles. Many teachers believe writing tests should be left to the professionals, a belief easily maintained since textbook publishers are quick to supply a list of questions accompanying every chapter. Professionals do construct the items: professional item writers, not professional educators. Many item writers try to emphasize main ideas and higher cognitive processes; some are more successful than others. Examine the test provided by the publisher covering a unit you have taught. How many items deal with the big ideas or ask the students to apply or analyze material?

Teachers as well as publishers sometimes succumb to writing quick, memory-level items. A study of 342 teacher-made tests revealed that most teachers use short answer tests measuring knowledge of terms, facts, and principles (Fleming & Chambers, 1983). The tests require students to remember but not apply knowledge. Such tests are easy to construct, but they send wrong signals to students about the things we value in education.

Fortunately, there is nothing mysterious about constructing better classroom tests. Three steps will lift your test out of the ordinary and provide a sounder basis for evaluating student achievement: test planning, item writing, and item analysis and revision. You will soon be reading a chapter on each of these steps, but an advance word may be in order.

Test planning begins with drafting a test blueprint or table of specifications designating the subtopics of the unit and the cognitive processes that will be required in answering the items. The cognitive processes are typically defined by the Bloom Taxonomy which emphasizes comprehension, application, synthesis, and analysis. The table of specifications is your best guide to writing tests that capture your teaching emphasis and reflect cognitive processes above the memory level.

Armed with the test blueprint, there is one and only one point at which tests become good or bad: writing the individual test items. Some items are so wordy and bulky that getting through the test is a matter of reading comprehension and endurance. Other items contain obviously wrong alternatives so that correct answers can readily be chosen by students who have only a hazy grasp of the material. Far too many items use direct phrasing from the textbook, thus giving a break to students who may recognize the correct answer without understanding it. The curatives for these ills in item writing lie in a few simple rules of item construction. Procedures for item analysis and item revision will round out your skills in constructing objective test items.

Although multiple-choice items can be constructed to measure comprehension and application, the essay test stands out as holding the greatest potentiality for measuring analysis and synthesis of conceptual material. The down side of essay tests is their susceptibility to many hazards in scoring. Scoring of essay tests, however, can be vastly improved by using some well-known precautions. While the research evidence is uneven, essay tests appear to encourage students to think about the issues and motifs in subject matter.

A new family of measurement procedures is arriving on the scene under the general term performance assessments. A popular book by Mitchell (1992) provides a good introduction to this new movement. Performance assessments include such tasks as essay writing and problem solving in mathematics and science. A distinguishing feature of performance assessments is scoring procedures directed at the processes employed by the learner. Ideally, the scoring yields insight into how the student approaches the problem and reveals snags and difficulties which can be remedied by subsequent instruction. While there are some continuing questions about the psychometric qualities of these approaches, i.e., the extent to which they meet the criteria for reliability, validity, and standardization procedures, performance assessments may have offsetting advantages in signaling clearly to students some of the more broadly conceived aims of instruction. This congruence between performance

testing and many of the educational reform goals has prompted the inclusion of performance tests in statewide testing in this country and in national testing programs abroad.

A related aspect of the general movement to restructure classroom practices is a renewed interest in the measurement of thinking. In our discussion of the measurement of thinking, we make a distinction between formal thinking, which focuses on such reasoning processes as induction and deduction, and “everyday thinking,” which centers around the way situations are perceived, organized, and interpreted. The existing tests of critical thinking grow out of the formal reasoning tradition, and most use a multiple-choice format. This text offers an approach to the measurement of thinking that rests on the assumptions of “everyday reasoning.” Scores depend on the way students interpret complex situations. The scoring rationale gives weight to developing the student’s own point of view, explaining rather than simply describing the situation, and considering many factors rather than making simple right-wrong judgements. Teachers can apply the scoring procedure to situations that are open to a variety of interpretations, such as taking sides on controversial issues.

The discussion of measuring thinking processes will complete the section on constructing classroom tests. One of the most important goals of any course in measurement is to help teachers write, analyze, and evaluate classroom tests. Tests are always written with an eye on larger educational goals, and an underlying message of the chapters on test construction is the importance of thinking about subject matter content. This orientation is maintained as attention ranges from the construction of multiple-choice items to the more complex task of assessing thinking processes directly.

Measuring Human Behavior

Teachers interested in the cognitive abilities, aptitudes, interests, and adjustment of their students will find a number of standardized tests in each of these areas. In many cases, well-established instruments dominate the market. For example, in the area of individual intelligence testing, the Stanford-Binet and the Wechsler tests are at the pinnacle of all tests available. These are the tests of choice when it comes to making difficult diagnostic decisions. Although the tests are administered by psychologists, teachers need to know about them since they receive the reports and recommendations.

Our introduction to this section starts with Alfred Binet, whose extensive work with his own daughters and with school children led to the development of the first intelligence test almost 100 years ago. The story