European University Studies

Dieter Thoma

Strategic Attention in Language Testing

Metacognition in a Yes/No Business English Vocabulary Test

PETER LANG

This book approaches the empirical validation of a popular method of assessing lexical knowledge – the yes/no vocabulary test or lexical decision task – from an interdisciplinary perspective. Each of its chapters sets out a distinct phase of the research along the development of a business English vocabulary test. The book offers a linguistic discussion of business English vocabulary, a psycholinguistic discussion of what aspects of linguistic knowledge and (meta)cognitive processing mechanisms are involved in completing the task, a psychometric-methodological discussion of how these aspects are measured and scored, and an applied discussion of the usefulness of the test. Thus, it is of interest to researchers, students, and professionals from a range of disciplines.

Dieter Thoma, born in 1976, is assistant professor of English Linguistics at the University of Mannheim (Germany). He studied at the University of Mannheim and the University of Swansea (UK) and received his Diploma (MSc) in Business, English, and Business Education and his Doctorate (Dr.) in English Linguistics and Psycholinguistics in 2003 and 2009.

www.peterlang.de

Strategic Attention in Language Testing

European University Studies
Europäische Hochschulschriften
Publications Universitaires Européennes

Series XXI
Linguistics

Reihe XXI Série XXI
Linguistik
Linguistique

Vol./Bd. 368

PETER LANG
Frankfurt am Main · Berlin · Bern · Bruxelles · New York · Oxford · Wien

Dieter Thoma

Strategic Attention in Language Testing

Metacognition in a Yes/No Business English Vocabulary Test

PETER LANG
Internationaler Verlag der Wissenschaften

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at http://dnb.d-nb.de.

Zugl.: Mannheim, Univ., Diss., 2009

D 180
ISSN 0721-3352
ISBN 978-3-631-60580-6 ISBN 978-3-653-01868-4 (eBook)

www.peterlang.de

Acknowledgements

First and foremost, I like to thank my first PhD supervisor Prof. Dr. Rosemarie Tracy. Rosemarie provided all the support and freedom I could ask for, irrepressible optimism, enthusiasm, and even more ideas.

I'm indebted to Prof. Dr. Paul Meara at the University of Wales in Swansea, UK. Paul gave me the initial inspiration to work on the yes/no vocabulary test and its scoring models years ago when I was a young exchange student in Swansea. He has paid continuant interest in my work ever since.

I would like to thank Prof. Dr. Hermann G. Ebner for becoming a supervisor at long and eventually short notice and for greatly supporting my research by awarding course credits to a large number of students who participated in my experiments.

This book wouldn't exist without the hundreds of students in Mannheim who volunteered to take my tests or who participated in my experiments. Special thanks to Christiane Koch, Prof. Dr. Dr. hc Alfred Kieser, and Prof. Dr. Sabine Matthäus for lending me their course time to administer my tests.

Thank you to Olaf Thiele who happened to sit in an adjacent office first and became a friend. Olaf provided me with all kinds of custom-made software, taught me about probability, we co-conducted our experiments, and he thankfully never gave up telling me that you can overdo everything.

Many thanks to my student research assistants who have accompanied me in my office over the years. Sebastian Frank, Anna Lüneberg, Ulla Spengler, and Sonja Withopf assisted me in data collection and coding, read drafts and endured my obsession for exact looking charts. You did all kinds of administrative work to save me some time and I think you actually enjoyed listening to my strange new ideas.

I owe many thanks to the people who know much more about maths than I do and who discussed my countless scoring models with me: Dr. Björn Böttcher, Katharina Hein, and Anne Thomas.

I very much like to thank my friends and colleagues who scarified their scarce time to read and comment on drafts of individual chapters or their parts: Dr. Ira Gawlitzek, Dr. Holger Hopp, Stephan Hartmann, Franziska Löhmann, Carolyn Ludwig, and Katja Tessmann.

A special word of thanks to my family for providing continuant support and a reliable down-to-earth alternative in the *Hotzenwald*.

Contents

List of figures

List of tables

Chapter 1 Introduction

1.1 What this book is about

As a consequence of increasingly competitive global markets, there is a growing need for language tests, in particular for tests of business English. Doing business in a foreign language without knowing the appropriate words is highly inefficient and can be financially dangerous. Language tests should enable their users to allocate human or material resources optimally. In terms of human resources, it is pointless to train somebody in a language they already speak well enough. In terms of material resources, it is unwise to hire applicants for a job for which a test indicated that their language skills are sufficient while they are actually not. Finally, in terms of economic resources, it would be wasteful to exclude a gifted applicant from university studies because of inappropriate language-based admission tests. The validity of inferences drawn from observed language test performance depends on how well we understand the cognitive processing mechanisms a specific test elicits.

On the most general level, this book seeks to answer the question of how much strategic attentional control humans have over their response behaviour in language tests. This question is important because all language tests require test takers to implicitly monitor and control their language performance, and because many language tests require explicit metacognitive judgements about linguistic knowledge. Successful test performance requires test taking strategies (Bachman 1990). In this context, a strategy is an adaptive mental programme that plans how the goal of test completion can be achieved. The crucial issue is the extent to which test takers can use their strategic attention to control their own test performance given what they know about and can do with a language for one thing, and independently of their linguistic skills, for another. Test developers should know in how far they can control for these strategies by design features.

This book investigates the issue of strategic attention in language testing by looking at the special case of a yes/no vocabulary test of business English. In a yes/no test, participants read vocabulary items presented on a checklist and decide for each item whether or not they know its meaning. It is a forced-choice task. The test format was invented by Anderson and Freebody (1982) and was transferred to the second language context by Meara and Buxton (1987). It contains a mix of words and possible yet nonexistent pseudowords. The words are a representative sample from a corpus-linguistically designed word list. The pseudowords are used to control for the reliability of the

responses because participants cannot possibly know what they mean. The standard assumptions for inferences based on yes/no test performance are that *yes* responses to words provide an estimate of vocabulary size, whereas *yes* responses to pseudowords can be used to adjust this estimate for guessing. A validation of a yes/no test needs to probe these causal assumptions. In particular, we need to know whether participants can attentionally control the cognitive processes that inform their decisions, or if they follow a response bias that forces them to respond rather *yes* or *no* independently of their actual knowledge. Given the focus on cognitive processing in test taking behaviour, the methodological relevance of this issue goes beyond the scope of language testing because it is a major concern in psycholinguistic word recognition experiments (see Balota & Chumbley 1984; Borowsky & Besner 2006).

1.2 Theoretical background and the aims of the study

Overall, this study represents a test validation conducted in the light of a modern validity theory by Shadish, Cook, and Campbell (2002), where validity is defined as the approximate truth of an inference. Validity is established by theoretical rationales and empirical evidence that explain variation in test performance and that are integrated in a coherent validity argument as proposed by Cronbach (1988), Kane (1992; 2006), and Mislevy (2003). This argument requires an understanding of both the language test materials and their interactions with mental representations in cognitive processing mechanisms that, in turn, lead to a particular test performance. This understanding needs to be implemented in evidence-based mathematical scoring models which provide the basis for inferences generalizing from test performance to language knowledge and to the ability to put this knowledge to use. Against the backdrop of these general considerations, the present book has five specific aims that can be outlined as follows.

The linguistic aim

The linguistic aim is to model a test construct of receptive business English vocabulary size based on both the rationales for vocabulary testing and those for the testing of languages for specific purposes (LSP). Vocabulary testing is a sub-discipline of language testing. However, vocabulary tests vary in their design and in their construct definitions for explaining test performance as the effect of the person, the test situation, or their interaction (Read & Chapelle 2001). Vocabulary testing is useful if it provides information about lexical knowledge and processing that informs linguistic hypothesis testing and educational decisions. I will argue that it is specialized vocabulary that makes the

difference between LSP tests and general language tests. Since vocabulary has the closest link to extra-linguistic content knowledge, a vocabulary-driven approach may contribute to the solution of the theory deficit generally associated with LSP testing (e.g. Douglas 2000; Davies 2001).

The psycholinguistic aim

The psycholinguistic aim is to build a model of yes/no vocabulary test behaviour on the grounds of theoretical and empirical findings on word recognition, on metacognitive decision processes, and on the interfaces between these domains. I advance a novel model, the Metacognitive Multiple Pathway Selection (MMPS) model which integrates a multiple pathway selection approach to lexical processing and lexical decisions (e.g. Stone & Van Orden 1993; Balota et al. 1999) with an attentional metacognitive decision framework (e.g. Nelson, T. & Narens 1994; Koriat 2007).

The empirical aim

The empirical aim is to conduct a validation study that provides evidence for the main hypothesis which holds that the yes/no Business English Vocabulary size test (BEV) permits valid inferences about the size of a learner's specialized receptive vocabulary of business English. The main hypothesis rests on four sub-hypotheses that are again operationalized in testable claims. The most fundamental of the sub-hypotheses posits that yes/no test decisions are metacognitive processes under strategic attentional control which draw on recall-based lexical knowledge. The remaining hypotheses test four central questions that can be raised on the basis of observed yes/no test performance. (a) Is the mere recognition of a sample of words a valid indicator of the size of a learner's receptive vocabulary? (b) If so, do the responses to words in the Business English Vocabulary test actually measure business English vocabulary size of second language learners? (c) Can the responses to pseudowords be used to make inferences about metacognitive decision and guessing behaviour? (d) If so, what are necessary and sufficient conditions that determine the boundaries between lexically informed decisions and blind guesses?

To test these hypotheses, a total of 720 participants took part in the validation study: 585 in paper-based tests and 135 in a computer-based laboratory study. All of them completed the main version of the BEV and some self-assessment tasks. Different sub-samples took part in one or several control studies. These were a yes/no vocabulary test reaction time experiment, a follow-up confidence rating scale and translation test, a test for morphological knowledge, and business English C-tests all specifically constructed for the purpose of this study, as

well as a test for general English vocabulary size by Meara (1996b), and a measure of risk attitude by Lejuez et al. (2002).

The psychometric-methodological aim

The psychometric-methodological aim is to create an adaptive mathematical scoring model that implements the response behaviour which a yes/no test elicits. Scoring can be understood as the bottleneck in testing because it draws together and consolidates the evidence gained from test performance, and it opens up the interpretative arena. Scoring thus lies at the interface between measurement and interpretation and bears substantially on validity. I will show that most of the statistical assumptions of existing yes/no test scoring models (e.g. Anderson & Freebody 1982; Meara 1992c; Beeckmans et al. 2001; Huibregtse et al. 2002) are not met by yes/no test response data. As an alternative, I propose a polynomial scoring model that makes fewer *a priori* assumptions about the statistical properties of yes/no test responses. Instead, the model is adaptable to the empirical distributions of the response data a particular yes/no test elicits.

The applied aim

The applied aim is the evaluation of the BEV with respect to its potential uses as a proficiency test, in placement decisions, and as a standardized means of self-assessment. As a yes/no test, the BEV is highly efficient and, subject to its validation, it might be able to meet the growing need (see O'Sullivan 2006) for theoretically and empirically substantiated business English tests.

1.3 How this book is organized

This book has eight chapters which are structured as illustrated schematically in fig. 1.1, where each chapter is represented by a box. The arrows between the boxes indicate how the chapters contribute to the five aims outlined above.

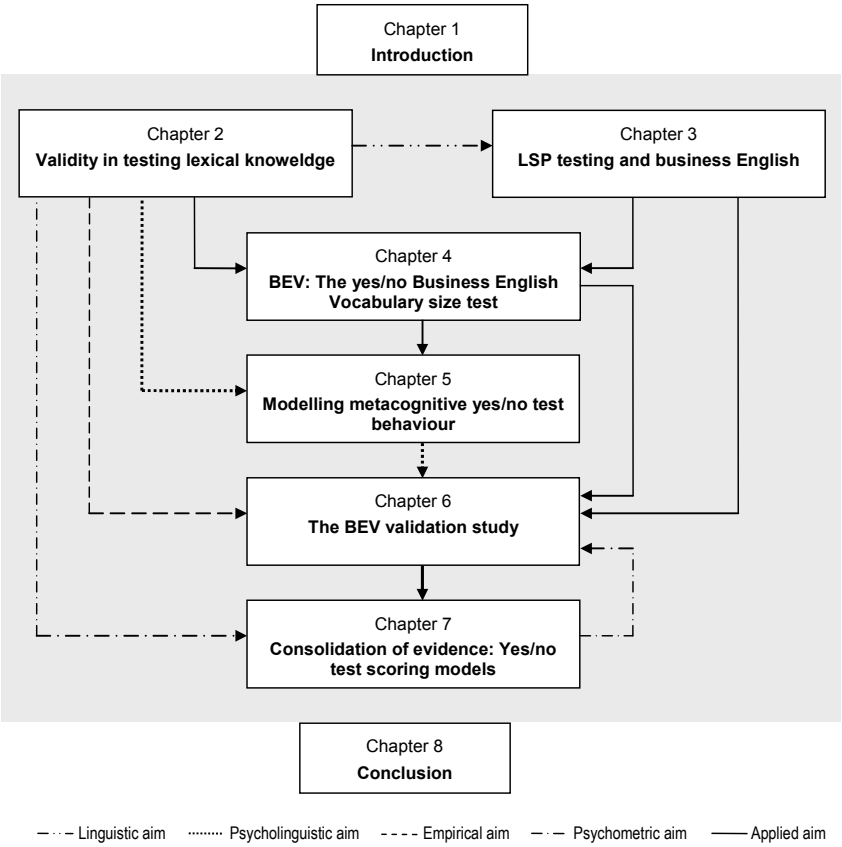

Fig. 1.1 Structure of this book

Following this introduction, chapter 2 first reviews the basic concepts of language testing and the evolution of validity theory. Second, it addresses the central questions of what vocabulary testing is about, how it can be operationalized, and why it is useful. It arrives at an understanding of validity, the effects of which are evident throughout all subsequent chapters.

Chapter 3 serves the linguistic aim of this book in that it develops a construct of business English vocabulary. To this end, the chapter explores the theoretical and empirical soundness of the two rationales brought forward in support of LSP testing: the existence of distinct language varieties, such as business English, and of distinct, context-dependent language abilities. The continuous arrows in fig. 1.1 indicate that the construct definition also serves the applied aim of this study in that it guides the test development in chapter 4, and in that it provides

the construct to which the inferences drawn in the validation study refer.

Chapter 4 has two functions. It first presents the yes/no test in detail and reviews the available studies on its measurement qualities. The results are summarized in a preliminary evaluation of the test format. The rest of the chapter presents an operationalization of the construct definition of receptive business English vocabulary size and gives a technical report of test development that led to the BEV. The arrows in fig. 1.1 show that the BEV is the product of the applied aim, it is subject to the validation study in chapter 6, and it provides the language test material with which the MMPS model in chapter 5 as well as the score functions in chapter 7 are evaluated.

Chapter 5 picks up where chapter 2 and chapter 3 left off, namely in their emphasis on vocabulary. In this chapter, I move towards the psycholinguistic aim of this book and build a model of metacognitive yes/no vocabulary test behaviour. For this reason, the chapter reviews the basic concepts and the major empirical findings in the areas of visual word recognition and metacognitive decision behaviour, on the one hand. On the other, it surveys existing models of lexical and metacognitive processing and evaluates their potential for explaining yes/no test behaviour. The chapter integrates these findings in the MMPS model, which is the theoretical backbone of the validation study.

Chapter 6 documents the BEV validation study and thereby realizes the empirical aim. The validation takes the form of a hypothesis testing procedure and presents a series of studies. As a by-product, the empirical evidence is used to test the predictions of the MMPS model. This largest chapter ends in a validity argument for the BEV. In addition, the chapter contains an evaluative summary on the test qualities of the BEV.

Chapter 7 explores statistical and mathematical scoring models for the yes/no test. Just as scoring is the bottleneck of test interpretation, this chapter is the bottleneck of this book. I evaluate existing models on theoretical criteria derived from the theory of yes/no test behaviour elaborated in the previous chapters and develop a novel model based on the grounds of this logic. I then evaluate all models on their empirical performance in transforming the test data observed in the BEV validation study into an informative score that allows for valid inferences about vocabulary size. The chapter thereby serves the psychometric aim. This book ends in a general conclusion in chapter 8.

1.4 How this book can be read

This book integrates a broad range of topics. However, dependent on the reader's interest, there are four basic options of how it can be read. The first option is, of course, to read the whole book in the order of its presentation. The second option is for readers with a special interest in the testing of business English vocabulary whom I recommend to read chapters 2 through 4 as well as sections 6.7 and 6.8. The third option is for those who want to focus on modelling and testing yes/no vocabulary test behaviour. They can selectively read chapter 2, chapters 4 and 5, and chapter 6, possibly without section 6.7. The fourth option is for readers with a special psychometric interest in yes/no test scoring models. They can read chapter 7 independently; maybe supplemented with some general information about the yes/no test from section 4.2. Moreover, chapter 2 holds a relatively self-contained discussion of validity theory in psychological and educational testing.

Chapter 2 Validity in testing lexical knowledge

2.1 Introduction

Research into language testing asks three fundamental questions: first, *how* linguistic knowledge and abilities can be measured, second, *what* it means to know and to be able to use a language, and third, *why* language tests are necessary. The most important theoretical concept in language testing is *validity* because it conjoins the linguistic question of *what* is being measured with the methodological question of *how* the measurement proceeds. It is highly controversial whether validity theory should also encompass the *why* question. To anticipate the result of the forthcoming review, I understand validity as the extent to which an inference based on a test performance is supported by empirical evidence and theoretical rationales as being true or correct. Validity is a property of inferences not of a test as such. One classic kind of evidence that supports validity is the consistency and thereby the accuracy of the measurement, i.e. reliability. Through its triple focus, illustrated in fig. 2.1, language testing research is an interdisciplinary area at the interfaces between linguistics, cognitive psychology, and psychometrics.

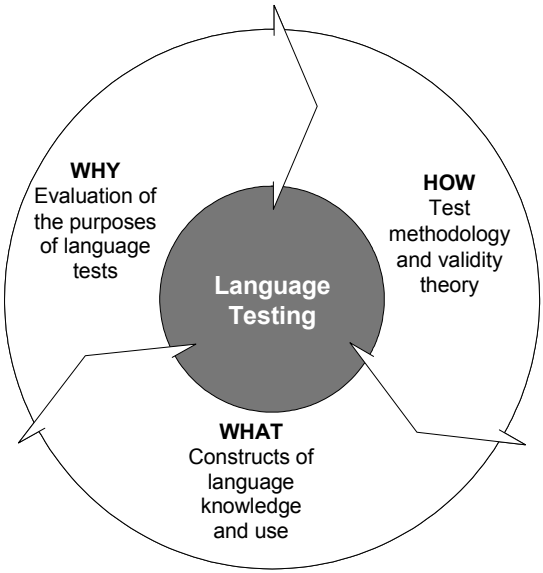

Fig. 2.1 The triple focus of language testing research

This chapter first outlines the key concepts and terminology associated with the three fundamental questions of language testing research. It is organized as follows. Section 2.2 presents definitions and theoretical approaches in the three core areas: (1) language testing methodology and psychometrics, (2) construct definitions of language knowledge and use, and (3) the evaluation of the purposes and uses of language tests.

Provided with this inventory, section 2.3 devotes considerable space to the review and discussion of the evolving, expanding, and controversial field of validity theory in general and in language testing in particular. The controversies in validity theory range over a historical dimension and over a current dimension. Both dimensions rest on the discussion of the scope of validity theory. Section 2.3.1 surveys the historical development of the theoretical approaches to validity in educational and psychological measurement. This historical review is necessary because there is no current consensus on the understanding of validity. Although validity theory has been evolving constantly, virtually all of the conceptualizations that were once brought forward have remained in use. The section starts with the origins of the development in the 1920s, when validity was conceived as one of several technical qualities of a test, moves on to a description of the traditional threefold conceptualization encompassing construct, criterion-oriented, and content validity along with reliability as test properties, and then documents the first turning point in the 1950s, when Cronbach and Meehl (1955) argued for a hypotheses-driven approach that focussed on the construct validity of inferences based on test behaviour. The section subsequently outlines the currently dominant view of validity as a unitary and far-reaching concept based on multiple sources of evidence coined by Messick (e.g. 1989) in educational and psychological testing and brought into language testing by Bachman (1990). This historical review ends in the description of recent argument-based validity frameworks which hinge on the idea of integrating empirical and theoretical evidence into a validity argument that follows the logic of Toulmin's (1958; 2003) argument scheme.

Section 2.3.7 presents an alternative view of validity understood as the approximate truth of inferences in the tradition of experimental psychology and coined by Campbell (e.g. Campbell & Fiske 1959; Cook & Campbell 1979; Shadish et al. 2002). Even though these frameworks share many parallels, this view presents one of the strongest current oppositions to the unitary view of validity in the tradition of Cronbach and Messick. Section 2.3.8 recapitulates the development of validity theory and shows that the conceptual differences between the two major traditions originated in the scope attributed to the concept of valid-

ity. While a unitary understanding of validity integrates validation and evaluation theory and practices, the alternative builds on the difference between truth and value and thus between validation and evaluation. The section ends in the discussion of possible implications of the controversial understanding of validity for language testing.

Section 2.4 turns to vocabulary testing as a sub-discipline of language testing. As such the testing of lexical knowledge is also characterized by the three fundamental questions of *how*, *what*, and *why*. The section therefore first describes how the object of measurement is conceptualized from a linguistic point of view and then addresses the question of how vocabulary can be defined as a test construct and how such a construct can be justified. If we assume that lexical knowledge is a cognitive module, there are two principal ways of approaching the object of measurement of a vocabulary test. First, vocabulary can be viewed from a corpus-based perspective as part of a language system, as part of a descriptive grammar. Second, knowledge of a vocabulary can be seen as a part of a mental lexicon that organizes the storage and the access to individual word representations in the mind. Both perspectives are necessary for the development of a vocabulary test. Sections 2.4.1 gives a brief survey of each. Detailed analyses follow in chapter 3 and chapter 5, respectively. Knowledge of vocabulary is multidimensional but the number of words known by an individual – vocabulary size – has been identified as the most important dimension for language learners. Section 2.4.3 therefore outlines some exemplary methods of standardized vocabulary testing with a special focus on measuring vocabulary size. Section 2.5 provides a summary of the key concepts in language and vocabulary testing.

2.2 The 'how', 'what', and 'why' in language testing

2.2.1 Language testing as a scientific method

The methodological *how*-question is based on psychological and educational measurement. There, the key term **test** refers to any *objective* and *standardized* method used to obtain empirical samples of human behaviour and includes the assignment of a (numerical) value to the elicited response according to a scale (Cronbach 1990; Anastasi 1997). The value that quantifies the observed behaviour is the test **score**. The test score should permit inferences that will generalize beyond the behaviour observed in the test context (Messick 1998a). The target of these generalizations are typically both the mental attributes that are supposed to cause the behaviour, called *constructs* or *traits*, and comparable behaviour in non-test situations. In language testing, the response behaviour observed in the test performance should permit ge-

neralizable inferences about what a test taker knows of a language and about the extent to which they can put this knowledge to use in comprehension and production.

The terms *test* or *testing* co-occur with *measurement* and *assessment*. **Measurement** is in some sense narrower than *testing* because it is (often) limited to the systematic quantification of behaviour, which is part of testing (see Stevens 1946 for the methodological origins). *Measurement* can, however, also refer to a broader concept, since tests are no necessary condition for the quantification of behaviour – naturalistic observations are but one alternative. There is also no definition clearly separating the terms *testing* and *assessment*. Although **assessment** is frequently used as superordinate covering all forms of quantitative and qualitative, standardized and individualized diagnostic procedures (e.g. Davies et al. 1999), it is also a regular synonym of *test* (e.g. Clapham 1997). As quasi-antonym of *testing* in the sense of *alternative assessment* it refers to relatively informal, non-standard, and small-scale methods based on alternative theoretical reasoning (Clapham 2000a). I mainly use word forms of *test*, particularly when concrete measurement instruments are in question.

Generalizations from test performances – including either broadening or narrowing inferences to more general or to more specific instances – are broadly made for two purposes: (1) in the process of hypothesis testing to investigate the language system itself or what speakers know about it and can do with it; (2) in educational settings to select and/or promote learners or to evaluate their teaching programs. Scores of educational tests can be interpreted in a norm-referenced or criterion-referenced manner (Bachman 1990; Kleber 1992). With a **norm-referenced** interpretation, the participants are ranked according to their relative test performance, i.e. socially with reference to others or individually with reference to their own previous performance. With a **criterion-referenced** interpretation, a test taker's performance is compared to a specified knowledge standard or level of ability. **Standardized** tests are measurement procedures that attempt to minimize the variation in content, administration, scoring, and interpretation across different versions of the 'same' test and its administrations by different testers and to different test takers (Gregory 1996: 33). Standardization requires empirical research and pretesting of the instrument (Bachman 1990: 74). When the standard is a social norm derived from the test performance of a relatively large participant sample against which an individual's performance is to be compared, the standardization process is referred to as **norming**. **Objectivity** more specifically refers to the extent to which test development, administration, scoring, and interpretation are independent of

the tester (Kleber 1992: 189ff). It used to be a major quality of a test but has lost importance because more recent approaches in test theory try to control for the effects of subjective factors rather than to exclude them (Bachman 1990; Suen 1990).

The systematic development of educational and psychological measurement procedures and the methodology of using tests to quantify psychophysical behaviour, abilities, and problems is the science behind psychological and educational testing and is referred to as psychometrics. **Psychometrics** develops (test) theories which provide the logical-statistical basis to answer the key questions of reliability and validity. There are two major branches of test or psychometric theories: Random sampling theory, which in turn comprises two approaches: classical test theory and generalizability theory, and item response theory. **Classical test theory** rests on the idea of a true score model where the *observed score*, the result of a test performance, is composed of the *true score* and the *error score.* The central aim of classical test theory is to improve the reliability of tests by closely estimating and minimizing the measurement error. Classical test theory is still very popular in psychological testing although it can only implement random sources of error and is unaware of systematic confounds. Moreover, it is strictly dependent on the (normal) distribution of the data provided by the participant sample (see e.g. Bachman 1990; Suen 1990; Nunnally & Bernstein 2006). **Generalizability theory** has been developed as an evolutionary alternative in the 1950s and implements reliability as a function of multiple sources of error. **Item response theory**, also called *latent trait* theory, was developed as an independent alternative in two lines of research headed by Birnbaum in the US and Rasch in Denmark (Davies et al. 1999). It is far less dependent on random sampling and is more suitable to predict individual test performance. While the true score model rests on the assumption that only a single trait or construct is measured by a test at a time (*unidimensionality*), item response theory can handle multidimensional constructs. The peculiarity of all test theories is that they cannot be confirmed empirically by the process of hypothesis testing but need to be verified by "logical deductions and mathematical proofs" (Suen 1990: 8). Otherwise the instrument would be method and result at the same time, which leads to irrational circularity.

A relatively recent methodological approach to test construction and validation are qualitative analyses that counter and complement more traditional quantitative and statistical psychometric methods (Bachman & Cohen 1998; Lazaraton 2008). According to Lazaraton's (2008: 198) summary, qualitative test analyses employ methods such as discourse/conversation analysis, questionnaires, observations, ver-

bal protocol analysis, and interviews. An exemplary study object of conversation analysis is the interviewer discourse in speaking tests. A typical feature of this kind of qualitative research is the "'unmotivated looking' at data rather than presenting research questions" (2008: 199).

2.2.2 Language as a test construct

The question of *what* is being tested in language testing concerns both the content of a test as well as the information it provides as a product. The result of a test is a performance sample that should allow generalizing inferences to a construct of relevant behaviour and thereby to similar behaviour in non-test situations. Language tests aim to obtain information on developmental stages of language learning and therefore most of them are used in the context of foreign or second language (L2) learning. The key concepts and problems do, however, not differ between first language (L1) and L2 testing. On the most aggregated level, the construct underlying the responses in a language test is attributed to **language ability**, or **proficiency** (Bachman 1990, 1998). Language ability is the conceptualization of what constitutes an individual's linguistic knowledge and their ability to use this knowledge comprehensively and productively. Tests are always imperfect measures because test performance is not only influenced by the construct under focus. Fig. 2.2 schematically illustrates that, in Bachman and Palmer's (1996) view, observed language test performance is the combined effect of language ability, the characteristics of the test method, and their interaction. Test performance is also affected by systematic and random confounds that again interact with the other causes. The construct serves a double purpose. It is (a) the basis for deriving representative instances of language behaviour and test methods, and (b) it is the reference point for generalizations, i.e. score interpretations from the observed test performance.

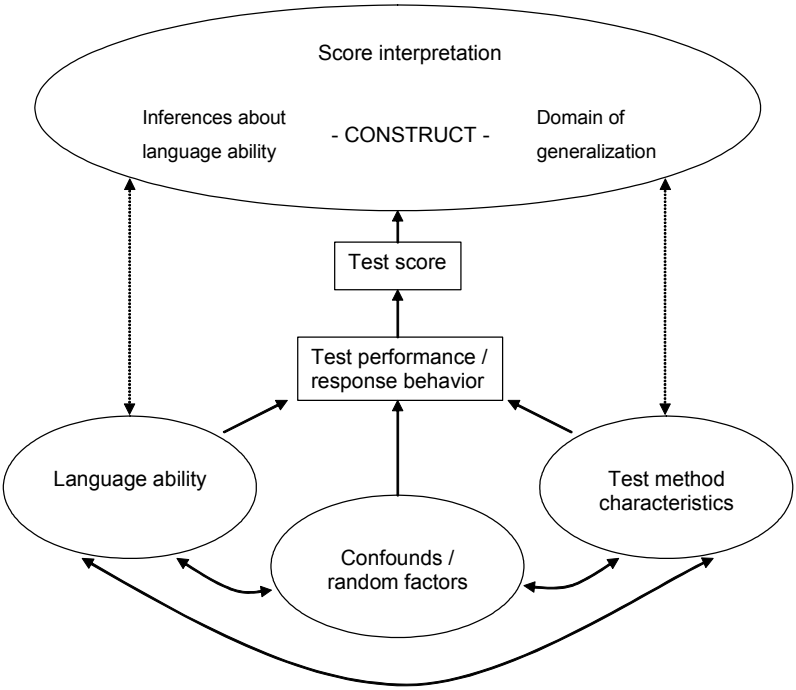

Fig. 2.2 Interpretation of language test performance as the combined effect of language ability, the test method, confounds, and the interaction of the independent variables, adapted from Bachman (1990: 165), Bachman and Palmer (1996: 22).

Language testing thereby follows the general question of causality in the social sciences, which addresses the extent to which observed behaviour is the effect of the person, the situation or both (e.g. Heckhausen 1989). From the perspective of construct definition, Messick (1989) elaborates on three approaches towards explaining consistent test performance: *Trait*, *behaviourist,* and *interactionalist* construct definitions focus on the person, the situational context, or the interaction between person/situation, respectively. Drawing on Messick's work, Chapelle (1998) applies these three approaches directly to language testing. In short, both authors describe them as follows:

According to a **trait definition**, response consistency is attributed to relatively stable characteristics of the test person. If language ability is defined as a trait, score variances can be mainly assigned to differences among individual participants, i.e. their individual proficiency levels. A **behaviourist definition** attributes consistent test performance to situational or contextual variables. In behaviourism, language as a mental ability is considered much too complex and obscure to be defined explicitly at all; it is a prototypical black-box phenomenon

(Skinner 1957). A behaviourist definition thus identifies situational factors which stimulate language performance. An **interactionalist definition** takes an intermediate perspective in that it attributes test performance to consistent behaviour that is the effect of traits, context, and the interaction between those two variables.

The understanding of language ability has changed hand in hand with the test methodology from one that was derived from a structuralist linguistics view during the 1960s and 1970s to a multicomponential, interactionalist view that culminated in recent approaches of communicative language testing (Bachman & Cohen 1998; Shohamy 2008 for an overview see tab. 2.1). Behaviourist definitions played at no time a significant role in large-scale testing. They are reflected to some extent in the developments alongside the major chronology such as *alternative assessment*. A behaviourist approach towards creating language tests would imply that no supposed language component or skill can be tested in isolation, and therefore only an "overall judgement of communicative effectiveness" (Read & Chapelle 2001: 8) is a valid score interpretation. The observed performance can only be generalized to contexts which are very similar to the test situation, i.e. if there is a very close correspondence between test setting and prospective contexts of language use. In language testing practice, however, such a view is often too far from measuring individual participant differences – which has always been one of the core purposes of the testing enterprise. Towards the end of the 1990s the social-political dimension of language testing has been discussed somewhat irrespective of linguistic theory and methodology in an approach called *critical language testing*.

Tab. 2.1 Historical interaction of *what* and *how* in language testing

Language testing theory	Construct	Method
Discrete-point testing	Structuralist view: lexical and grammatical items	'Objective' tests with many decontextualized items
Integrative testing	Skills and components	Skills tests: reading, writing, listening, speaking; cloze tests
Communicative testing	Communicative language ability	Performance tests, tasks in simulated 'real life'
Alternative assessment	Language knowledge is extremely complex	Multiple and various procedures
Ethical and critical language testing	Educational and social-political consequences of testing	

The relation between the construct of language ability and language testing methodology is to some extent a catch-22 situation. On the one hand, valid inferences from test scores require the elicited response behaviour to be a representative and consistent sample of the construct. On the other, test methods are one major way to gain knowledge about the nature of language ability. For this theoretical circularity, particular design features of tests are closely related to the conceptualization of the underlying construct. And particular design features are unlikely to provide fully conflicting evidence about their preset construct. The differences between methodological test types are therefore in practice the extreme points on a continuum.

Discrete-point tests (discrete tests for short) set out to test lists of decontextualized linguistic items or components – mainly grammar and vocabulary – or more precisely, a selected sample of them. On a slightly more integrated level, the focus is on language skills, typically reading, writing, listening, and speaking. Test performance is attributed to a trait definition of language ability. Discrete tests built on traditional psychometrics with a major emphasis on the improvement of reliability. Since reliability is a function of the number of measurement points, it is best achieved with many small test items. In addition, discrete language tests attempt to present their many items in isolation to diminish context effects because these are seen to be the main sources of measurement error. To meet the assumptions of psychometric random sampling theories, discrete language tests rely on random samples drawn from the relevant domain of items that is intended to represent the construct. Individual test items are considered independent of each other. The idea of discrete tests is that if discrete components of language are measured, it is possible to draw inferences that will generalize beyond the particular test items, because these are a representative sample of the ability modules making up language use in non-test situations. The best-known example of discrete language tests are multiple-choice tests that pose a question and offer several response alternatives out of which at least one is considered the correct response. Discrete tests had their heydays in the psychometric-structuralist period in the 1960s when the construct of language ability was seen as an aggregation of highly decomposable units (see Lado 1961). The main disadvantage in discrete point testing is seen in the assumption that the knowledge of the units of a language equals the ability to use the language in authentic settings. Discrete tests measure only declarative language knowledge whereas the inferences drawn from language tests are mainly used for decisions that require information about a learner's ability to communicate effectively in the language. Effective communication requires declarative and procedural

knowledge. Despite the criticism, discrete point tests "have remained hugely influential" (McNamara 2000: 14) – presumably because of their strength in quantifying language behaviour and because language knowledge is not a sufficient but a necessary condition for effective communication.

Integrative tests are essentially multi-skills tests based on the mediating interactionalist approach to construct definition. The theoretical underpinnings of integrative language tests specify personal traits, contextual features, as well as metacognitive strategies used to apply language traits in communicative situations (Chapelle 1998: 9). Lewkowicz (1997) proposes to distinguish between two extreme poles along a continuum of multi-skills tests. Tests at one pole simply require participants to use more than one language skill at a time, e.g. a dictation task involving listening and writing, whereas fully integrative tests are performance-based in that they require the test taker to complete a certain communicative task. Examples of integrative tests are oral interviews to assess speaking and written compositions. The disadvantages of such methods are that they are less economic and potentially unreliable as compared to discrete tests. So-called reduced redundancy procedures such as the cloze test (Oller & Conrad 1971) and C-test (Klein-Braley & Raatz 1984) that require test takers to complete deletions in words, phrases, or texts have been proposed to overcome these problems. Empirically, however, reduced redundancy procedures are similar to discrete tests and often provide similar results (McNamara 2000). **Performance tests** are at the other extreme of the continuum as they attempt to maximize the contextualization of language ability. McNamara (1997) identifies two directions in performance testing: the *work-sample approach* intends to replicate the communicative demands of real-life situations with a focus on the quality of the language (e.g. English for academic or for occupational purposes); the *cognitive/psycholinguistic approach* wants to mirror "the contextualised on-line processing required in all performance" (p. 131), e.g. the interaction between specific language skills and subject knowledge required for writing an academic essay.

Communicative language testing uses performance tests and, unlike integrative approaches, explicitly attends to the social roles of the test takers (McNamara 1995, 2000). Similar to strongly integrative tests, it rests on Hymes' (1972) construct of **communicative competence** or later extensions of this theory. Hymes' original *communicative competence* was first intended as a counter-concept to Chomsky's (1965: 4) distinction between *competence* (abstract language knowledge) and *performance* (actual language comprehension and production). Chomsky's (1965) focus is on knowledge of language relatively

independent of specific social contexts of communication. Competence models describe grammatical knowledge about the constituents of a language and combinatory rules. Performance models describe pragmatic competence in language processing. Both models are only accessible via performance data displayed in actual language use.

Hymes' (1972) model still contains one component he calls "knowledge" and another one he calls "ability for use", but both are part of the abstract model of communicative competence. He thus attempts in particular to integrate sociolinguistic factors of language use into Chomsky's predominantly psychological model (see McNamara 1995 for discussion). The standard model of communicative competence in applied linguistics is by Canale and Swain (1980). Following this and other previous approaches Bachman (1990) developed a model of **communicative language ability**, which has been the most influential construct in language testing to date. Similar to psychometric-structuralist approaches, Bachman's communicative language ability comprises components and skills, but the structuralist conceptualization has been extended. It recognizes (a) that there is broader range of linguistic components, (b) that these interact, (c) that language ability interacts with extra-linguistic abilities, and (d) that language ability includes interaction with the communicative context (Bachman & Cohen 1998: 6). The first three points can still be attributed to personal traits, which exist independently of any situational factors, but the fourth requires the implementation of the situational context in a theory of language ability. Accordingly, Bachman (1990) as well as Bachman and Palmer (1996) define *communicative language ability* as being composed of declarative components embedded in *knowledge of language*, and procedural components that constitute *strategic competence*. Communicative language ability interacts with extra-linguistic general cognition ("knowledge structures") and is mediated by psychophysiological mechanisms, which are "involved in the actual execution of language as a physical phenomenon" (p. 84). Bachman's model is depicted in fig. 2.3. The fundamental distinction between actual performance data and abstract knowledge is less sharp than in Chomsky's and Hymes' models, mainly because the model is bound to explaining test performance.

Knowledge of language (language competence) consists of the integration of a relatively large number of components (see fig. 2.4). Formal aspects of language such as vocabulary, morphology, and syntax are subsumed under "organizational knowledge"; functional aspects such as manipulative functions of language and sensitivity to register under "pragmatic knowledge".

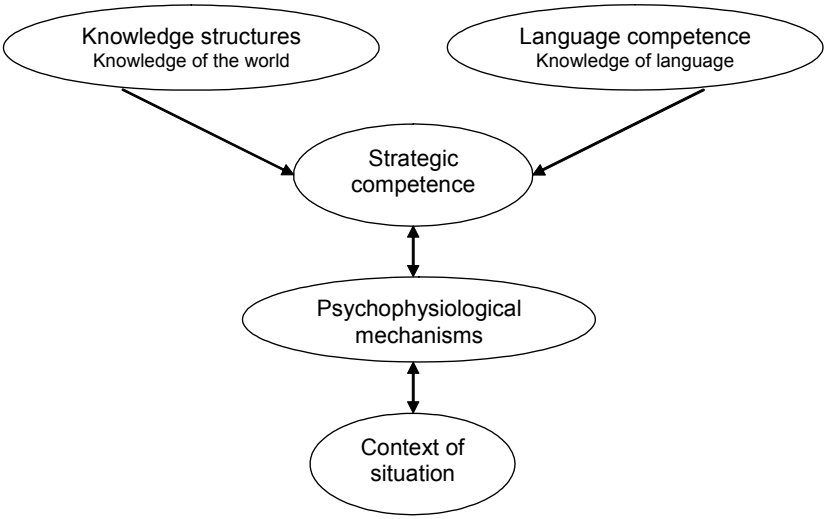

Fig. 2.3 Components of communicative language ability in communicative language use, adapted from Bachman (1990: 85).

The implementation of functional aspects into the construct abilities enables the consideration of the test taker's social role. *Strategic competence* is "the mental capacity for implementing the components of language competence in contextualized communicative language use" (Bachman 1990: 84). Bachman and Palmer (1996: 70) describe strategic competence as a set of metacognitive strategies that integrate language knowledge, topical or subject knowledge, and affective schemata in online language use. The metacognitive strategies are part of the construct of communicative language ability and thus also the target of generalizations from the test performance. They are, however, also specific strategies selected to master the test task. Bachman and Palmer (1996: 71) identify three areas of metacognitive strategy use: goal setting, assessment, and planning. *Goal setting* involves the decision whether or not to complete the language task or test the test taker has identified and selected. *Assessment* involves the assessment of the resources required to complete the task and the monitoring of the correctness of the own responses. *Planning* involves the selection and planning of the integration of topical and language knowledge as well as the selection of one plan to produce a response.

The advantage of Bachman and Palmer's model is that by its grounding on language skills and components it provides the theoretical basis for tests that are more informative than pure performance tests based on holistic, behaviourist constructs of language ability in a strict tradition of communicative language testing.

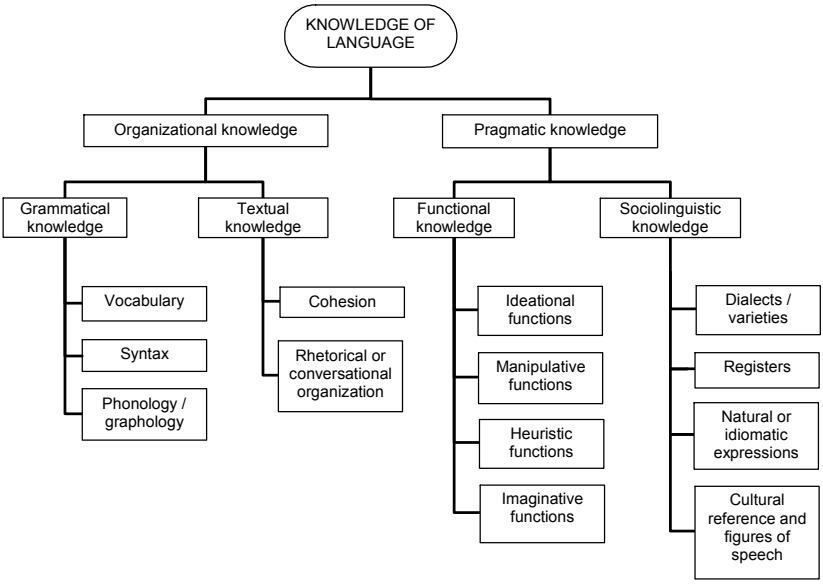

Fig. 2.4 Components of language knowledge; adapted from Bachman (1990: 87) and Bachman & Palmer (1996)

Inferences about the overall communicative effectiveness of a language learner may be sufficient for general placement decisions, but tests undertaken for educational or research purposes often require more specific information about language knowledge and use.

To return to the chronological development of language testing, every new form of assessment is **alternative assessment** at first so that task-based performance tests were also covered by the term at times. The dominant feature of alternative assessments is, however, that they focus on the pedagogical assessment of individual learners rather than on standardized proficiency testing. Alternative assessment pays more attention to qualitative methods than to the quantification of human behaviour. The field has some proximity to behaviourist definitions of the construct of language ability given that alternative assessment considers language knowledge as an extremely complex phenomenon that can only be probed by multiple and varied procedures (Shohamy 2008). Chapelle and Brindley (2002) list informal observation of learner language, learner portfolios that document individual progress, and self-assessment as the major methods of alternative assessment. One pedagogical motivation is, for example, that learners trained in self-assessment become aware of relevant assessment criteria and are enabled to optimize their learning methods – a

component of strategic competence in Bachman and Palmer's (1996) model. Self-assessment thus in principle serves diagnostic purposes. Fully uncontrolled or unverified self-assessment is certainly inappropriate for selective purposes such as grading or certification (Oscarson 1997).

The most recent developments in language testing emphasize the socio-political role of tests. McNamara (2006) identifies *ethical* language testing and *critical* language testing as two major paradigms. **Ethical language testing** (e.g. Alderson & Wall 1993; Davies 1997; Hamp-Lyons 1997; Lynch 1997) is concerned with the improvement of the ethics and fairness of the test process itself by evaluating and setting professional standards for test developers and test users as well as by considering the direct consequences tests have on language learning and teaching. **Critical language testing** is a direction founded by Shohamy (1998; 2001) that concerns itself with "the need to question the uses of tests as tools of power and to examine their uses in education and society" (2001: 376). Critical language testing (see also Lynch 2001) has been motivated by Messick's (1989) extension of validity theory to consequential aspects of testing (see next section) and Pennycock's (e.g. 1994; 2008) focus on the social-political dimension of international English language teaching, and it runs in the tradition of radical social theorists such as Fouclaut and Bourdieu. Bourdieu (1984), for example, posits that educational certification systems serve social reproduction and reserve elite positions. Since language proficiency tests often regulate the admission to higher educational systems, job markets, or even citizenship, their gatekeeping function can be seen as integral part of this social mechanism (e.g. McNamara & Shohamy 2008). In sum, critical language testing widens the horizon of the field of language testing by adding a critical *why* to the questions of *what* and *how,* but it neither has the intention nor the scope to contribute much to the understanding of the two original questions.

2.2.3 Purposes of language testing

The generalizability of test results entails a logical problem that is reflected in the dichotomy of discrete and integrative tests and in the related distinction between *direct* and *indirect* tests. Empirical samples of behaviour are by definition only snapshots of observable behaviour shown by eventually very specific individuals in specific situations, who engaged in very specific activities or received very specific treatments. These snapshots are taken with very specific equipment, i.e. specific methods of measurement or tests. One could argue that performance tests are more direct tests because the test tasks closely mimic the task that is the target of the generalization (McNamara

1996). Fig. 2.5 schematically illustrates that **indirect** (and discrete) tests require inferences from test performance via the construct to performance in non-test situations. **Direct** (performance) tests attempt to skip the construct stage and permit generalizing inferences from test performance to performance in non-test situations directly. From the perspective of psychometric validity theory, truly direct tests are hardly possible because, given that there is empirically no perfect match between two situations, the relation between test performance and inferences is logically always indirect (Bachman 1990). The logical problem rests in the degree of correspondence between the test and the non-test situation, the authenticity of the test task.

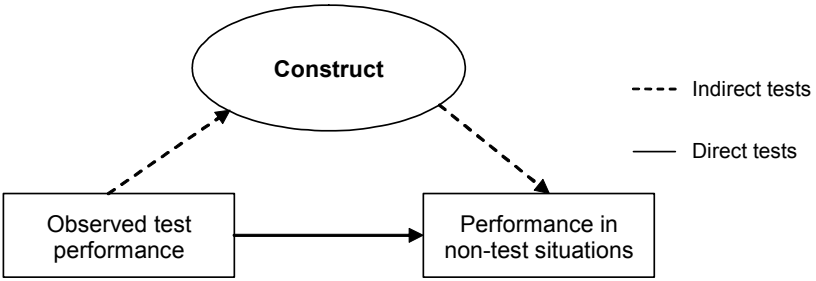

Fig. 2.5 Hypothesized routes for inferences based on test performance in direct and indirect tests

The issue of **authenticity** has become very popular during the 1990s in educational testing (e.g. Wiggins 1992) and has always been important in L2 acquisition research (e.g. Mitchell & Myles 1998: 3; Tarone 1998). The basic problem is as old as testing itself: To what extent do test tasks need to correspond to non-test situations, and what makes a sample relevant and representative so that valid generalizations about a participant's ability can be made from the limited test performance? These questions are associated with two dilemmas. The first is called the *observer's paradox* (Labov 1972). The very act of observing human behaviour may change it. Standardized language tests elicit response behaviour that in part only occurs in test situations, e.g. specific test completion strategies or phenomena that have been called test wiseness or test anxiety (Bachman & Palmer 1996; Davies et al. 1999). This dilemma can only be reasonably solved for practical purposes if language testers are aware of their interferences and consider them explicitly when they interpret test performance.

The second dilemma is referred to as *bandwidth-fidelity dilemma* in random sampling theory. Cronbach (1990: 208) puts it as follows: "Should the tester go after a good measure of one dimension or make several less thorough measurements?" In psychometric terms, the bet-

ter the fidelity (i.e. reliability and objectivity) of a test, the more constricted is the generalizability of the inferences from the test to comparable non-test situations, since contextual (erroneous) effects are minimized. However, an authentic, highly contextualized test sample, which considers multiple variables, allows testers to draw more generalizable but less reliable and objective inferences. The odd thing about this dilemma is that it can be converted to argue in favour of the complete opposite if the conceptualization of validity is changed. When the test performance is very authentic and considers multiple context-specific variables, the tester may be able to make inferences about very specific, similar non-test situations, but may not be able to generalize them to slightly different real-life situations, for the very reason of their authenticity (see Reusser & Stebler 1999 for discussion). In contrast, discrete tests may enable the tester to make generalizations about language performance beyond the test context just because they measure context-independent, multifunctional modules.

The solution to the bandwidth-fidelity dilemma – and thus to the question of how much trait, context or interaction is actually necessary for construct definition and valid score interpretations – is the determination and reasonable justification of the degree of approximating authentic language use in tests. These justifications cannot be founded on purely theoretical grounds. Language testers need to produce sufficient and integrated empirical and logically derived evidence for their uses of the test and its purposes (Messick 1989; Chapelle 1998). Whether these justifications will be generally accepted is yet another problem.

This leads directly to the purposes of language testing. How a language test is developed and interpreted crucially depends on its intended purpose. Language tests can be designed for research or educational purposes or both (Read & Chapelle 2001). The obvious purpose of research tests is to either verify or falsify hypotheses about language (as a system) or language ability (as a person's declarative and procedural knowledge about the system) by evidence from empirical data. In education, selecting persons and modifying persons or situations (i.e. promotion or intervention) can be seen as extreme points along a continuum of functions (Kleber 1992). Tab. 2.2 gives an overview of language test types and their mostly educational purposes.

Diagnostic and achievement tests are frequently so-called teacher-made tests that have not been developed and validated extensively in accordance with formal procedures. If teachers are trained in test construction and administration such tests can thoroughly meet their objectives.

Tab. 2.2 Test types and purposes in educational language testing (Brindley 2001; Nation 2001)

Test type	Test purpose
Placement (classification)	Knowledge assessment and assignment to a class or level.
Achievement	Assessment of short/long-term results of learning and teaching.
Diagnostic	Knowledge assessment and learner motivation.
Evaluation (accountability)	Improvement of teaching and programmes, evidence for educational authorities to justify expenditure.
Proficiency	Particular focus on learner's ability to apply language, often outside the learning context; gatekeeping.

Placement, proficiency, and evaluation tests, in contrast, are usually standardized tests based on psychometric theory and validation. Tests that proved to be practical, efficient, and readily available seem to cause the most intended and unintended cross-impacts between education and research. For example, Nation's (1990) Vocabulary Levels Test was originally intended as an aid for classroom teachers to plan and encourage systematic and learner-adequate vocabulary learning/teaching. Over the years, several researchers have adopted it as a research measure of vocabulary size (Schmitt et al. 2001). Standardized proficiency testing is the main business of the test industry. The importance of proficiency tests is mainly due to their gatekeeping function. The performance on such tests is used as a criterion in decisions about the admission of nonnative speakers to academic programmes and to jobs. Tab. 2.3 shows examples of tests produced by American and British organizations that test proficiency in English as a Foreign Language (EFL).

Tab. 2.3 Professional English proficiency testing organizations and their tests

Language testing organization		Well-known EFL test (example)	
ETS	Educational Testing Service	TOEFL	Test of English as a Foreign Language
ETS		TOEIC	Test of English for International Communication
IELTS	International English Language Testing System/ (Service)	IELTS	International English Language Testing System
UCLES	University of Cambridge Local Examinations Syndicate	FCE	Cambridge First Certificate in English
UCLES		BULATS	Business Language Testing Service
	University of Cambridge ESOL Examinations	BEC	Business English Certificates

The purpose of a test can have a direct influence on the definition of the construct of language ability and its components. Language can be used for specific purposes, e.g. for communication in academic, business, or medical settings, or for very general day-to-day purposes. Most tests of specific language abilities are performance-based tests in the tradition of communicative language testing. Chapter 3 addresses the theoretical justifications and the empirical problems of such tests extensively. In practice, test specificity is a matter of degree. On the one hand, Language for Specific Purposes (LSP) tests face a constant competition from general purpose language tests such as TOEFL or IELTS for the reasons of generalizability discussed above. On the other, LSP tests have some intuitive appeal to testers and test takers for the authenticity of their tasks. They also constitute a promising growth market as shown by recently developed standardized commercial tests for business language (e.g. BEC, BULATS). The books by Douglas (2000) on the assessment of LSP in general and by O'Sullivan (2006) on testing business English both survey and categorize existing LSP tests. They thereby become unintended examples of the finding that the applied purposes of language tests often outrun the questions of *what* and *how* in language testing. In other words, first there is a test which is used for some purpose, and only later there are endeavours to find a theory that justifies test contents, methods, and use. Validity theory is diametrically opposed to such practices.

2.2.4 Section summary: Language testing

This section covered the basic concepts of (1) language testing methodology and psychometrics, (2) construct definitions of language knowledge and use, and (3) the evaluation of the purposes and uses of language tests.

A language test is a standardized and objective method used to obtain and quantify empirical samples language knowledge and use. The test score should permit generalizations to the underlying construct of language ability and to performance in comparable non-test situations. *Measurement* is a more technical, *assessment* a more general term than *testing*. Standardized tests minimize variation in content, administration, scoring, and interpretation. The standard can be a norm or criterion reference. In a norming procedure, the standard is inferred from the performance of a large participant sample on a specific test. Objective tests minimize the influence of the tester. Psychometrics is the theory of testing. Its major branches are classical test theory, generalizability theory, and item response theory.

The most abstract target of generalization in language testing is the construct of language ability or proficiency. Its understanding is coined

by Bachman's construct of communicative language ability (1990; Bachman & Palmer 1996) which consists of knowledge of language, extra-linguistic world knowledge, and strategic competence. These factors interact via psychophysical mechanisms with the context of the communicative situation. Language test performance is explained as the effect of language ability, test method characteristics, random factors, and their interaction. The construct serves the double duty of allowing for generalizations by giving meaning to the test score and of providing the basis for the development of test items.

More generally, test performance can be explained as the effect of personality traits, (trait definition), situations (behaviourist definition), or their interaction (interactionalist definition). Strict behaviourist construct definitions cannot be aligned with standardized language testing because they cannot explain interindividual differences. Language testing started from a psychometric-structuralist approach that utilized discrete-point tests with decontextualized vocabulary and grammar items. These tasks refer to a trait construct, are objective, and can produce generalizable results, but they are not very authentic. By expanding the items toward reading, writing, listening, and speaking skills language testing became more integrative. Communicative language testing refers to an interactionalist construct of communicative language ability, which is operationalized in performance tests that try to approximate authentic 'real-live' communication. These three major historical developments are supplemented by alternative assessment approaches that favour behaviourist construct definitions and a more pedagogical emphasis on qualitative methods. Ethical and critical language testing addresses the educational and socio-political consequences of tests.

Indirect tests assume a construct mediates inferences from test performance to performance in non-test situations. Direct tests should permit generalizations to non-test performance directly. For this purpose, direct tests need to be very authentic in that they closely resemble the target situation. Truly direct tests are psychometrically not possible because there is no perfect match between two situations and there are task effects (observer's paradox). There is a tradeoff between authenticity and generalizability because highly authentic measurement permits very limited generalizations (bandwidth fidelity dilemma) whose extent can only be justified by the purposes of language testing. Language tests have research and educational purposes. Major educational purposes are placement, achievement, diagnostic, accountability, and proficiency testing. From the perspective of the purposes of the language material, there are tests of language for general purposes and language for specific purposes, e.g. business English.

2.3 Validity in language testing

2.3.1 Approaches to validity

"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed in proposed uses of tests", according to the latest version of the joint test standards of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999: 9). This definition represents an intermediate position between the most comprehensive unitary validity conceptualization in the tradition of Cronbach (e.g. Cronbach & Meehl 1955; Cronbach 1971, 1988) and Messick (e.g. 1989; 1995) and less far-reaching approaches (e.g. Cook & Campbell 1979; Wiley 1991; Popham 1997; Shadish et al. 2002). The current standard validity theory in language testing is coined by the work of Messick (Bachman 2005; McNamara 2006). In a seminal chapter on validity, he defines it as follows:

> Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores and other modes of assessment. (Messick 1989: 13)

The review in the subsequent sections will show that Messick's unitary conceptualization of validity as an integrated evaluative judgement capitalizes on an evolutionary history of validity theory. Current theories of validity are only understandable in the light of the history of validity theory in educational/psychological measurement and in language testing. Sections 2.3.2 to 2.3.6 provide this historical background, but they do not represent a strictly chronological review of the history of validity theory (for such reviews see Messick 1989; Geisinger 1992; Chapelle 1999; Kane 2004). While it can be argued that validity theory has undergone an "evolution" (Sireci 2007) or "metamorphosis" (Geisinger 1992), none of the conceptualizations of validity that has been brought forward so far seems to be totally extinct. Older and newer views coexist in research practice and in theoretical reasoning.

Messick's approach is still the most comprehensive and challenging (McNamara 2006; Xi 2008). The unitary and all-inclusive concept of validity has, however, been criticized for being impractical or even counter-productive (e.g. Popham 1997; Borsboom et al. 2004; Lissitz & Samuelsen 2007) and for blurring important boundaries between validation and evaluation (e.g. Wiley 1991; Shadish et al. 2002). Shadish, Cook and Campbell prefer to limit the scope of validity to the truth of inferences, which they separate from the evaluation of tests and testing:

> We use the term *validity* to refer to the approximate truth of an inference. When we say that something is valid, we make a judgement about the extent to which relevant evidence supports that inference as being true or correct. (Shadish et al. 2002: 34)

Section 2.3.7 outlines the Campbellian approach to validity. In essence, the two major paradigms in validity theory share many features but differ in their scope (see fig. 2.7). Campbell's understanding of validity is narrower and operates with validity-as-truth. Messick's (and to some extent Cronbach's) notion of validity reaches further in that it operates with validity-as-value. Both paradigms, however, establish validity as a property of inferences, of propositions or knowledge claims, of interpretations of observed test performance. The forthcoming discussion will show that the classical understanding of validity as a property of a test itself, which has recently experienced a theoretical revival (e.g. Borsboom et al. 2004; Lissitz & Samuelsen 2007), suffers from a logical error.

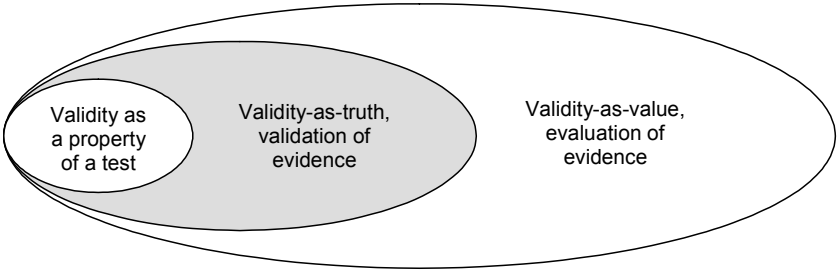

Fig. 2.6 Scopes of validity in different theoretical paradigms

Validity is so essential in testing that it touches fundamental philosophical concepts such as truth, value, and justice. Test **validation**, the process of providing validity evidence, is the core of language testing research and the methods used in validation are derived from and justified by a theory of validity (Xi 2008). At the same time *validity* is a heavy-duty word in the testing literature so that its exact meaning depends on the writer, the time of writing, and the research context. As a consequence, there is an overwhelming number of 'x+validity' terms. They range from completely atheoretical notions such as *face* validity – basically if a test seems about right for its purpose if a tester or a test taker just 'look' at it – via highly abstract concepts such as *construct* validity – addressing the match between theory and empirical data in its narrow sense (Cronbach & Meehl 1955) – to a notion of validity constituting an overall evaluative judgement of how well evidence and theory support certain score interpretations and test uses as in the selected definitions above. The proliferation of validity labels also has

been spurred by an urge of some researchers to replace established, although theory-laden, terms with their own coinages that often seem to differ little in their reference from meaning extensions of classical ones. An example are Weir's (2005) terms *theory-based* validity for *construct* validity and *context* validity for *content* validity, which even his own student O'Sullivan (2006) glosses constantly.

The conceptualization of validity in language testing has been largely conform to that in educational measurement in general. However, since large-scale language tests are the most international concern of testing research as well as extraordinarily big business, the field partly goes by its own rules. It often has taken key figures like Bachman (1990; 2005) to introduce novel concepts of validity into the field, and the language testing community seems quite resistant against fundamental criticism from adjacent research areas. For example, many articles in a recent compendium on language testing and assessment (Shohamy & Hornberger 2008) do not take account of the criticism brought forward against the scope of Cronbach's and in particular of Messick's validity theory.

2.3.2 Origins: Validity as a property of tests

Since its emergence around the 1920s, the concept of validity has continuously gained importance in psychological and educational measurement and in language testing. So for example, whereas validity used to be discussed somewhere in the last third of earlier textbooks (e.g. Lado 1961: chap. 23; McNamara 2000: chap. 5), it now often prominently resides in the very first content chapter (e.g. Weir 2005; Fulcher & Davidson 2007). The earliest definitions in educational measurement identified validity as a property of a test that is fulfilled if a test measures what it is supposed to measure:

> A test is valid when it measures what it purports to measure. (Kelley 1927: 14)
>
> The validity of a test is the extent to which it measures what it purports to measure. (Garrett 1937: 324)
>
> Does a test measure what it is supposed to measure? If it does, it is valid. (Lado 1961: 321)

This definition is still popular in psychological, educational, and applied linguistic contexts and is still given as a standard (e.g. Domino & Domino 2006; Daller et al. 2007; Schnitzer 2008). The experimental psychologist Hull (1928) operationalized the concept of test validity, although without explicitly defining it, as the extent to which the test score correlates with a test of free recall of the same material. In the wake of the rise of large-scale standardized psychological tests, such as