2008-07-05

The German Proficiency Exam at Brigham Young University: A Validation Study

Tina Grahovac Starr
Brigham Young University - Provo

THE GERMAN PROFICIENCY EXAM AT BRIGHAM YOUNG UNIVERSITY: A VALIDATION STUDY

by

Tina G. Starr

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Arts

Center for Language Studies

Brigham Young University

August 2008

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Tina G. Starr

This thesis has been read by each member of the following graduate committee and by majority vote had been found to be satisfactory.

Date	David K. Hart, Chair
Date	Diane Strong-Krause
Date	Michelle S. James

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Tina G. Starr in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

David K. Hart
Chair, Graduate Committee

Accepted for the Department

Ray T. Clifford
Director, Center for Language Studies

Accepted for the College

Joseph D. Parry
Associate Dean, College of Humanities

ABSTRACT

THE GERMAN PROFICIENCY EXAM AT
BRIGHAM YOUNG UNIVERSITY:
A VALIDATION STUDY

Tina G. Starr
Center for Language Studies
Master of Arts

In order to continuously improve the teaching and learning in a language program, it is a crucial part of program evaluation to assure that its assessment instruments have a beneficial influence on the teaching and learning procedures. For that reason, evidence was gathered to investigate the validity of test scores of the German Proficiency Exam (GPE) used by the German Section of the Germanic and Slavic Languages Department at Brigham Young University.

The GPE consists of seven exam components: listening comprehension, reading, writing, speaking, grammar, vocabulary, and strong verbs. The GPE component scores of 179 students were used to conduct the analysis for this study. In order to estimate the reliability of the test scores, Cronbach's Alpha was calculated

for the listening comprehension exam, the reading exam, the grammar exam, the strong verbs exam, and the vocabulary exam. In addition, the analysis included overall descriptive statistics, item facility and item discrimination, distractor analysis, ANOVA, and a post-hoc Tukey's pairwise comparison.

The results of the Cronbach's alpha indicated relatively high reliability of scores of all the exam components except the listening component. The item and distractor analysis of the strong verbs and vocabulary exam revealed that the scoring procedures need to be revised so that the scores reflect a student's true knowledge. The descriptive statistics of the exams showed a limited usage of the scoring range and suggest defining the scoring procedures and training the scorers. Further, it was suggested to define a general language construct and the specific construct of each language skill on the basis of which proficiency levels can be developed. Using the results of the data analysis various suggestions were given to improve the validity of scores of the GPE.

ACKNOWLEDGMENTS

This work would have not been possible without the encouragement and the immense support of my friend, mentor, and committee member Dr. Diane Strong-Krause. With her positive attitude, her inspirations, and her great efforts in spending a lot of time to explain things clearly and simply, she helped to make research more enjoyable for me. Throughout my thesis-writing period, she provided encouragement, sound advice, good teaching, and a lot of good ideas. I would have been lost without her.

I would also like to thank my thesis chair Dr. David Hart and my committee member Dr. Michelle James for taking the time to work with me during this endeavor.

I wish to thank Agnes Welch in the Language Studies department who has been a constant support during the five years of my graduate school helping me to be on track with deadlines, courses and credit hours and to receive the scholarships making graduate school possible.

I would also like to extend my gratitude to the faculty and staff of the German department. In the seven years at BYU they have always been willing to help with any kind of matter. They have provided encouragement to finish graduate school, and they have believed in me.

I especially would like to thank Dr. Hans-Wilhelm Kelling who made it possible in the first place to come to BYU. He is a great example and like a grandfather to me.

I cannot finish without saying how grateful I am for my family. My parents, Nada and Zeljko Grahovac, have raised me to always strive for knowledge, progress continuously and become a better person. They have sacrificed so much for me and my siblings to have a good life, receive good education and to have a loving and safe home. My family has been a great blessing and support for me throughout my life and especially during my hardest times.

I can't find any suitable words to express enough gratitude for my husband Nathan. I love him without end. To him I dedicate this thesis.

Lastly, and most importantly, I would like to express gratitude to my loving Heavenly Father who has heard my innumerable prayers and blessed me with many helpful people. He has given me energy, motivation, guidance, and many miracles so that I did not give up and was able to write my thesis, a miracle in itself.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

Introduction

Rationale for This Study

"When any scholar is able to read Tully or such like classical Latin author ex tempore and make and speake true Latin in verse and prose, suo Marte, and decline perfectly the paradigms of nouns and verbs in ye Greeke tongue, then may hee bee admitted into ye college, nor shall any claim admission before such qualifications." (Harvard College, 1642)

One cannot ignore that foreign language testing is an integral part of the teaching-learning process. As constituted in the admission standards of the Harvard College, someone had to determine whether an applying student really had "such qualifications." Thus, foreign language testing has been with us for several hundred years. In the last hundred years, since Cambridge's Certificate of Proficiency in English (CPE) was first offered, and the 'scientific' issue of test reliability was still relatively little understood (Weir, 2005, 5), language testing has made tremendous progress. The question of *what* and *how* we are testing has been and still is a critically discussed topic in the field of education.

Brigham Young University (BYU) strives to continually improve the learning of its students. In his address at the annual university conference faculty session in 2005, John S. Tanner, academic vice president at Brigham Young University observed that "a serious institutional commitment to lifelong learning … has profound implications for how we teach our students. It forces us to focus less on what we teach and more on what they learn. This can be a difficult paradigm shift for those of us who sometimes indulge exclusively in the "sage-on-the-stage" model of

teaching. It is, however, a paradigm shift that for more than a decade has radically altered the landscape of higher education." To implement this paradigm shift mentioned by Tanner (2006) and also to fulfill the requirements for reaccreditation the Northwest Commission has requested that BYU

- Identify and publish expected learning outcomes for each of its degree programs;
- Demonstrate that students who complete their programs have achieved the stated outcomes; and
- Provide evidence consistently across its programs that its assessment activities lead to improvement of teaching and learning (Tanner, 2006).

The German Section of the Germanic and Slavic Languages Department, henceforth 'German Section', at Brigham Young University followed this request and has identified and published expected learning outcomes for each of its degree programs. Furthermore, in order to demonstrate that students have achieved these stated outcomes, the German Section has established direct and indirect measurement methods, or assessment activities, to show evidence of learning. One of the direct measurement tools is the German Proficiency Exam (Germanic & Slavic Languages, 2007), which all German majors are required to take in connection with the German course 400R during one of the last two semesters before graduation. The purpose of the German Proficiency Exam (GPE) is to determine how well students graduating from the German program perform in the skill areas of listening, reading, writing, speaking, grammar, and vocabulary. As mentioned above, the Northwest Commission has also requested that each program provide evidence that its assessment activities lead to improvement of teaching and learning. In other words,

the German Section is required to show that the measurement tools have a beneficial influence on the teaching and learning procedures, which is also known as a positive washback effect.

For this reason the faculty and staff of the department want to know whether the GPE measures what it is supposed to measure. In particular, they want to examine the validity and reliability of the GPE. This research study outlines an initial investigation into the validity of the GPE. Its purpose is to answer the question: To what degree is this test valid and reliable? Hughes, Porter, and Weir (1988) underline the need for validity evidence: "The provision of satisfactory evidence of validity is indisputably necessary for any serious test" (p. 4). If a test is not valid, it might be questionable whether the test scores are an accurate representation of a student's level of language knowledge or skills, and decisions that are made on the basis of these scores are founded on shaky grounds.

According to Kunnan (1998), since the 1960's the main focus of language testing has been validation. In the recent years many language testing researchers have adapted Messick's (1989a) view on validity which he defines in this way: "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). In a validation study the goal is to gather various types of evidence, to evaluate to what degree an instrument measures what it is supposed to measure and to what degree the meaning and interpretation of scores are properly used for decision making. In order to make a sound evaluation, it is not sufficient to make a claim of validity on

the basis of only one single type of evidence, but to collect numerous forms and aspects of evidence to determine to what degree the test scores are valid. In the field of language testing one or another form of evidence has been identified to support 'types of validity.' But because all of these types of evidence constitute validity, validity is seen as a unitary concept (Messick, 1989b).

Since validity is such a complex concept, and the process of validation is very extensive, it is not possible within the timeframe of this research study to conduct a complete validation study; instead, this study concentrates on specific relevant questions that will help the German Section.

Purpose of This Study

This research study gathers validity-related evidence to help answer some questions concerning the validity of the German Proficiency Exam at Brigham Young University. The German Section is interested in improving the quality of the GPE, and this study collects qualitative evidence and quantitative data from test scores, analyzes them, and attempts to interpret the quantitative analysis to make suggestions for improvement, which will later be used to positively influence the teaching and assessment process for future students of the German program.

In the following chapter, relevant literature in the field of language testing and program evaluation are reviewed and discussed. On the basis of this review and discussion of literature, specific research questions are formed that function as guidance for this research study. In chapter three, I describe the research methodology used to gather qualitative information and quantitative data as evidence to help answer the research questions. The subjects involved in this study

are described, the measurement instrument – the German Proficiency Exam – presented, and the procedures for conducting this study explained. The results are presented in chapter four. In chapter five, the implications of those results are presented and discussed. Based on the findings from the qualitative and quantitative data, conclusions are drawn regarding the validity of the GPE. This is followed by suggestions for the improvement of the German Proficiency Exam.

Operational Definitions

1. *Language proficiency* is a term that has usually been used in the field of language testing "to refer in general to knowledge, competence, or ability in the use of a language, irrespective of how, where, or under what conditions it has been acquired" (Bachman, 1990, p. 16). The term 'communicative competence' which also refers to language ability, is however used in a broader sense than language proficiency (Bachman, 1990). In this study, I generally prefer to use the term 'language ability,' however, sometimes the term 'language proficiency' or 'communicative competence' are used interchangeably with 'language ability.'
2. Proficiency scale (also rating scale): "A scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged" (Davies, 1999, p. 153). A proficiency (rating) scale provides a definition of a construct such as proficiency. The levels of a proficiency (rating) scale are usually defined by what subjects can do with the language and their ability in the various language skills and features.

Limitations

As mentioned earlier, due to the limited timeframe, an extensive and in-depth validation study is beyond the scope of this research study. Instead, this project focuses on answering specific questions regarding the validity of the GPE scores. As such, the following aspects are limitations of this study:

This study does not address the administration of the exam. The conditions under which the GPE is administered to students will not be considered.

Although it would be helpful in a validation study of this nature, this research does not analyze whether the content covers all aspects of each of the skill areas.

There is some investigation of the general construct coverage of this exam. This study does not investigate whether the tasks or items of each skill area cover the full construct of each skill area.

CHAPTER TWO

Review of the Literature

Chapter one outlined the importance of and need for a validity study of the German Proficiency Exam (GPE) of the German Section at Brigham Young University (BYU). The concept of validity was briefly mentioned, and the need for gathering validity-related evidence to help verify the effectiveness of a given test was explained.

The purpose of chapter two is to provide a theoretical basis for conducting such a research study. This chapter addresses the approaches of program evaluation and describes the role of testing in the process of program evaluation and improvement. Different test designs and their functions are discussed. Also, language ability is explained and a definition of a language construct is given. The details of and the need for a validity study are described. Finally, the connection between validity and reliability is established, reliability is defined and the importance of reliability is described.

The Role and Use of Testing in Program Evaluation

In chapter one it was explained that the Northwest Commission has requested that each degree program at BYU (1) identify and publish expected learning outcomes; (2) demonstrate that students who complete their programs have achieved the stated outcomes; and (3) provide evidence consistently across its programs that its assessment activities lead to improvement of teaching and learning (Tanner, 2006), not only to fulfill the requirements for reaccreditation, but especially to achieve the improvement of learning at BYU. Each degree program essentially is

asked to perform three basic steps in program evaluation. Notice that the third step requests evidence for the validity of tests. Before discussing the meaning of validity, it is important to understand the process of program evaluation and its necessity. Additionally, on must understand the role testing plays in the broad picture of program evaluation and how different test designs can be used in program evaluation. The purpose is to clarify that testing is not an isolated part in the process of teaching and learning, but that it is closely connected with all the elements of a program.

Approaches and Dimensions of Program Evaluation

As was stressed in BYU's report to the accreditation committee (BYU, 2006), in order to meet the mission of the university and the aim of a BYU education, "consequently, the design, implementation, evaluation, and continual improvement of quality programs in recognized fields of study that lead to valid degrees are institutional priorities" (p. 2.2). It is crucial that programs of any sort undergo continuous evaluation to ensure that the teaching and especially the learning processes are meeting the goals. Beyond simply meeting goals, the goals or objectives must ensure that an effective learning process is taking place. In his article about program evaluation, Brown (1989) gave a general definition of evaluation as "the systematic collection and analysis of all relevant information necessary to promote the improvement of a curriculum, and assess its effectiveness and efficiency, as well as the participants' attitudes within the context of the particular institutions involved" (p. 223). This definition points out that it is not only sufficient to collect relevant information. The information must be collected systematically and

thoroughly analyzed. Further, there are two purposes described for the collection and analysis of information: to improve the curriculum, and to evaluate the effectiveness of the curriculum. Finally, this definition points out that the assessment must be channeled towards a specific curriculum for a program that is directly influenced by the institutions connected to the program. These influences can be a university administration, accreditation commissions, or the prospective employers (school districts, for example).

Over the years, different approaches to program evaluation have emerged which, according to Brown (1989), can generally be grouped into four categories: product oriented approaches, static characteristic approaches, process-oriented approaches, and decision facilitation approaches.

Product-oriented approaches have the purpose of determining whether the goals and instructional objectives of a program have been achieved. Thus, a program should be based on clearly defined goals and measurable behavioral characteristics, such as students, the subject matter, societal considerations, philosophy of education and learning philosophy. At the end of the program, the objectives should be measured and successful achievement of objectives determined.

The *static approach* to evaluation typically involves a group of outside experts who determine whether a program is effective or not. Usually, this sort of evaluation is connected to an institutional accreditation process, like the one BYU underwent in 2006. For such an evaluation, the institution is forced to provide any records relevant to the effectiveness of the program and demonstrate the adequacy of the physical learning facilities. The expert group doing the evaluation assesses in detail the quality

of the program based on the information described above in order to formulate a report based on their observation.

While achieving program objectives is very important, it is critical to understand that evaluation procedures can also be used to help change and improve the curriculum. This understanding leads to the adoption of *process-oriented approaches*. Some of the most important foci of program evaluations are (1) the distinction between formative and summative evaluation; (2) the importance of evaluating not only whether the goals have been met but also whether the goals themselves are worthwhile; and (3) goal free evaluation, i.e., the evaluators should not only limit themselves to studying the expected goals of the program, but also consider the possibility that there were unexpected outcomes which should be recognized and studied (Brown, 1989, p. 226).

Finally, in the *decision facilitation approach,* information is gathered for those who make judgments about and decisions for the program. These are usually the program administrators. Information is collected to help make decisions about the state of the overall system, program planning, program implementation, program improvement, and the overall value of the program.

Programs conducting an evaluation usually draw on several or all of these approaches. Depending on the circumstances of the program and the type of decisions that need to be made, the evaluation can take place in different "dimensions" (Brown, 1989). These dimensions are closely connected with each other. Each dimension is comprised of two perspectives, both of which should be

considered in an evaluation, inasmuch as each perspective can provide valuable information.

The first dimension is comprised of the *formative and summative perspective*. The purpose of a formative evaluation is to improve the teaching and learning process of a program and takes place during the development of a program. The information gathered gives insights in the results of the program, its strengths and weaknesses. A summative evaluation takes place after the completion of a program and helps in making the decision of whether a program is successful and effective, and whether a new curriculum should be adopted.

The two views of the second dimension are *process or product oriented*. A product-oriented evaluation is concerned with whether the program goals and objectives are being achieved. In process-oriented evaluation, information is gathered that gives feedback about the procedures used to arrive at the goals.

Finally, the last dimension has the *qualitative and quantitative perspective*. The difference between formative and summative evaluation lies in the purpose for which information is gathered, while the difference between product and process-oriented evaluation lies in what information might be considered. Alternatively, the difference between qualitative and quantitative evaluation lies in what type of information is being collected. Quantitative information is basically data that can easily be turned into numbers and statistics, such as test scores, student rankings, number of students in a class, etc. Qualitative data, on the other hand, are generally observations that cannot be turned into numbers or statistics easily. Qualitative data might include interviews, classroom observations, or journal entries (Brown, 1989; Hudson, 1989).

There are many procedures available for collecting the information essential for evaluating where and whether elements of a particular program need to be changed (Brown, 1989, p. 233). Some of these procedures help with gathering quantitative data and some with qualitative information, as was explained previously. Although a wide array of procedures are available to the evaluators, after reviewing the literature on program evaluation, it seems that testing is being used primarily and is most frequently discussed.

It is very beneficial in a program evaluation to be aware of the approaches and dimensions described above, so that the evaluator can choose the most appropriate type of information and the best way to examine that information. This awareness can help in making decisions for the improvement of the program. For a thorough and effective curriculum, it is important to gather as much information as possible covering as many perspectives as seem necessary in an ongoing evaluation and improvement of the program.

Language Curriculum Components

At the English Language Institute (ELI) at the University of Hawaii, the ongoing evaluation is an integral part of its curriculum. Figure 2.1 depicts the working model for curriculum evaluation adapted by the ELI and shows how testing interrelates with other elements in the process of evaluation.

Brown (1989), a professor at the University of Hawaii, explains that "in a systematic approach to curriculum design such as this, the primary information-gathering and organizational elements include the needs analysis, instructional objectives and testing. The information and insights gained from these activities can

then be analyzed and synthesized in the design of materials and delivery of instructions" (p. 234). With a cursory glance at the model implemented at the ELI, someone might think that all five steps need to be followed one after another, but that would only be the most ideal case in first developing a curriculum. However, in most cases a program is already running and well entrenched when the evaluation progress begins. In a situation like this, all five steps may occur at the same time.

For a well-grounded program, evaluation and improvement of these elements should be an ongoing process that binds the elements into a whole. These elements must be part of a comprehensive evaluation. Without the process of ongoing evaluation, any or all of the elements may become meaningless.

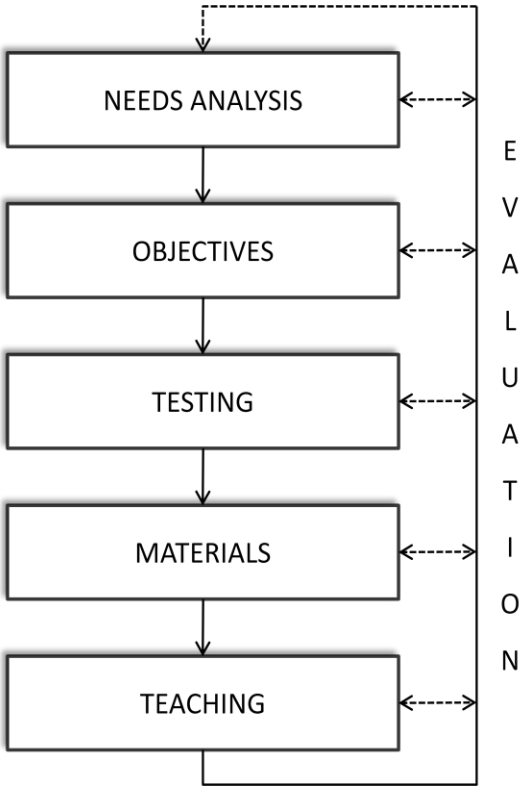


Figure 2.1. Systematic approach for designing and developing language curriculum (Brown, 1989, p. 235).

Test Designs

As mentioned earlier, Brown (1989) lists 24 procedures that can be used to gather information for an evaluation. Four of these procedures fall under the broad category of testing: placement, diagnostic, achievement, and proficiency (p. 233). Each of these types of test can provide different information.

A *placement test* has the purpose of measuring a student's language ability in order to place the student in a certain level appropriate for his or her ability. Usually, placement tests are used to assign students to classes at different levels. A placement test can either be based on a theory of language proficiency or on the learning objectives of the program.

Diagnostic tests, as their name suggests, are used to identify the strengths and weaknesses of students. They give insight in what students know or do not know about a language, or how well they can or cannot master the language skills. The information gained from diagnostic tests can be used to know in what areas further instruction is necessary.

Achievement tests are based on the syllabus and assess what the students have been taught or what they have learned over a specific period of time. Achievement tests are directly related to language courses, the course syllabus, textbook, or other materials used in instruction. Palmer (1991) indicates that the advantage of an achievement test is that it does not have to defend the course objectives. It only has to show that it covers a reasonable sample of the materials taught in the course. This advantage, on the other hand, can be a major disadvantage if the syllabus is poorly designed and the materials are badly chosen. In this case, the results of such a test

would be misleading, and even if students perform well on the test, it may not mean that they have met the course objectives (Hughes, 2003). Hughes suggests basing the test on the course objectives, which would compel course designers to identify course objectives more clearly. Further, it would also put more emphasis on using course objectives to create a well aligned syllabus and choose relevant course materials.

Proficiency tests, on the other hand, are not based on the content or objectives of language courses, but they are based on a theory of language ability the program chooses to follow. Thus, a proficiency test measures the student's ability in the language according to the specifications the language department has set as to what should be considered 'proficient.' That means that the student has sufficient competence in the language for a specific purpose. For example, a test can be designed to see whether someone can perform well in speaking the language in a business setting. Another example could be to find out whether a prospective student can function well in the language as a university student. However, a proficiency test does not necessarily have to have a specific setting or occupation in mind. It can be based on a more general theory of language proficiency and be based on 'program-neutral' goals. Palmer (1991) explains that the advantage of proficiency tests lies in being based upon a program-neutral theory of language, which enables it to measure whether a program reaches program-neutral goals. This demonstrates the importance for test developers to identify clearly what theory of language ability the test is based on and to develop the test using the methods of testing that support the stated theory. Using proficiency tests requires test developers to ask two major questions: "What is the nature of the language competence, and what evidence do we have that the tests

we are using actually measure that competence" (p. 3)? These two questions illustrate the basic concerns about the German Proficiency Exam. What indications are there that the German Proficiency exam is accurately measuring the proficiency in German of the graduating German students at BYU? Before going into more detail about these indications, language ability and construct definition are examined more closely. Without understanding language ability and construct definition, it is not possible to answer the question of whether a proficiency test is measuring what it is supposed to.

Language Ability and Construct Definition

Bachman and Palmer (1996) point out that it is necessary to define language ability clearly and in detail in order to set it apart from other individual characteristics that can affect test performance, such as test method (multiple choice, cloze, translation, etc.). In addition, a precise definition of language ability is essential to making conclusions about an individual's language proficiency on the basis of performance on a language test. Language ability should be defined in a way that is appropriate for the testing situation or the specific purpose, which becomes the basis for the conclusions that are made from the test performance. For example, when a patient requires surgery, he or she can be confident that the surgeon is able to perform the surgery safely, because the surgeon has been licensed only after demonstrating a high level of proficiency in the various skills needed for surgery by passing a series of examinations administered by the licensing body. Previous to developing this licensing examination, the licensing body clearly defined what the skills are that a surgeon needs to know and be able to do. Regarding the field of

language testing, Bachman and Palmer (1996) explain: "When we define ability this way, for purposes of measurement, we are defining a … 'construct'. In designing, developing, and using language tests, we can define our construct from a number of perspectives, including everything from the content of a particular part of a language course to a theoretical model of language ability" (p. 66). In order to make inferences or base decisions on test scores, one must make sure that the test is measuring the identified construct, or what it is supposed to measure. The concept of construct validity is discussed and described in the sections that follow.

Knowing what purposes these various test designs have and having a clearly defined concept of the language ability as the basis for a language test can contribute immensely to the development of a test that can effectively assess what it is intended to assess. In his study, Palmer (1991) conducted research looking into test design and research design issues. He compared and analyzed eight studies that used different language test designs in program evaluation. Palmer found that in most cases, the tests were classified by language use skill, sometimes by language ability, and sometimes by method. A description of the theory of language abilities upon which the tests were based was missing in all cases. And no consistent distinctions were made between language ability and test method. Palmer came to the conclusion that there is a trend of deficiency for tests in the following areas: "They lag behind recent work in language testing research; they use tests which are based upon models different from those that the methods' developers had in mind when they developed their methods; and they use tests which tend to avoid the issue of the distinction between language trait and testing method" (p. 6). This conclusion reflects the

situation for most language tests that have been developed. Tests are rarely based on a clearly defined construct and are therefore using testing methods that are not appropriate for the situation. This can cause the test scores to be unreliable. The reason for this, according to the deficiencies described by Palmer, is that the test developers do not inform themselves about the current theories of language testing and do not apply those theories in the design and development of tests.

Norm-reference v. Criterion-referenced Testing

In order to interpret test scores, a frame of reference is necessary. The two most common types of testing are the norm-referenced (NR) and criterion-referenced (CR) tests. *Norm-referenced* test scores are interpreted in reference to the performance of a group or norm, meaning that conclusions about performance on a test are made by comparing the individual student's score to the scores of the rest of the group. The 'norm group' usually is a large group of individuals for whom the test is designed. For example, imagine that a speaking test is given to an individual student. To know how this student performed on the test, one can say that she obtained a score that placed her in the top five percent of students that have taken the test, or the bottom ten percent. Another way to rank performance is to indicate if the student did worse or better than the average of the whole group (Bachman & Palmer, 1996; Hughes, 2003). The performance characteristics commonly used as reference points are the mean, or the average score of the group, and the standard deviation, which indicates how spread out the scores of the group are. If a test is designed well, the scores of a norm-reference test will usually be distributed in the shape of a bell-curve (Bachman, 1990). Very often, the test results of a student are interpreted and reported only in

reference to the group that took the test at that time. This is also called 'grading on the curve,' where, for example, the top ten percent receive an 'A' and the bottom ten percent fail, no matter how much their absolute knowledge of the material covered by the test was.

The strength of a norm-referenced test, and also the reason why this type of test is so widely used, is that it is easy to develop. Since in NR tests students are compared to other students, and not to levels of ability, the nature of language ability does not necessarily have to be defined. Even though this makes the test easy to construct, it creates a major weakness: test developers can avoid clearly defining the language ability they are testing for. This implies another weakness of NR tests. They do not provide a measure of how much of the language someone knows, and thus do not give the kind of information that would suggest what level of language ability was reached.

This weakness on the other hand is the strength of *criterion-referenced* tests. In criterion-referenced tests (CR) the scores of an individual are not compared in reference to a group, but in reference to pre-determined criteria that are independent of the way the other students scored on the test. The advantages of criterion-referenced tests are that an individual's score is not compared with that of other candidates, but rather describes what that individual can actually do in the language. There are two ways a CR test score can be interpreted (Bachman, 2004). Pophan (as quoted in Brown & Hudson, 2002) defines this first approach, which is sometimes called 'domain-referenced' or 'objectives-reference,' "as any test that is primarily designed to describe the performances of examinees in terms of the amount that they

know of a specific domain of knowledge or set of objectives. At any rate, the key factor is having a clearly described assessment domain" (p. 5). This approach is often used with achievement tests, where the domain of content is specified by the syllabus. An individual's performance on the test is determined by how much of the domain of content and objectives were mastered.

The other approach is where the language proficiency is described on a continuous scale, from no proficiency at all to perfect proficiency. "An individual's proficiency at a given task falls at some point on the continuum, as measured by behaviors he displays during testing. The degree to which his proficiency resembles desired performance at any specified level is assessed by *criterion-referenced measures of proficiency*" (Glaser and Klaus as quoted in Bachman, 2004, p. 31). One such rating scale most commonly used with language tests in the United States is the ACTFL set of proficiency guidelines. A student's language proficiency level in the skills of listening, speaking, reading, and writing can be described as novice, intermediate, advanced, superior or distinguished. For each of these skills, all the levels have specifications on what an individual has to know to reach that level. In Europe the Common European Framework is commonly used in language tests to describe the proficiency of a test taker. The levels of proficiency range from A1 to C2 (Council of Europe, 2004).

Being aware of the different types of tests, how they can be used in various circumstances, and what kind of information the scores provide can help in order to make correct decisions about what test method is most appropriate for the specific test situation. However, it is not sufficient just to be aware of the kind of information

tests can provide, it is necessary also to ensure the quality of information tests give. The decisions made in educational programs usually are about people, and the decisions affect their lives in one form or another. For that reason, it is essential that the information the decisions are based on is as reliable and as valid as possible. The following section discusses the issue of validity and reliability.

Validity and Reliability of Test Scores

Test Score Validity

Validity is the degree to which reliable test scores are interpreted correctly and used appropriately for making decisions. In the past, validity was considered a characteristic of a test. Lado (1961) defined validity this way: "Does a test measure what it is supposed to measure? If it does, it is valid" (p. 321). This view, however, is very limited since it only concentrates on the test itself and does not consider the interpretation and use of the test. Messick (1989b) emphasized that the validity of a test can never be justified if the interpretation and use of the test are not accounted for. Thus, validity should be regarded as the summation "of both the existing evidence for and the actual as well as potential consequences of score interpretation and use" (p. 5), and not as a characteristic of the instrument itself. Messick's (1989a) publication of his seminal paper was a very influential event in the field of educational research and language test development. He defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 13). As mentioned earlier, the decisions made in educational programs on the basis of an individual's

test scores affect their life in one form or another. Using test scores to put decisions into action always carries social consequences. Referring back to the example used earlier about a surgeon, it would be very tragic if a hospital hired a surgeon on the basis of high scores received on a set of required exams, but the surgeon turned out to be incapable of performing safe surgeries. The social consequences might include the death of a patient, which would have far-reaching consequences for the patient's family, the surgeon's family and the hospital. Or, for example, if a school district hires a foreign language teacher on the basis of high scores on a language proficiency test, the school district expects the teacher to be proficient in the language skills. If the teacher couldn't speak, understand, read or write the language well, he or she would not be an effective teacher. To justify an action based on test scores requires not only the validation of score meaning but also of value implications and action outcomes.

It is important to note that validity is a matter of degree and not an absolute condition. Furthermore, over time new findings of evidence will supplement the existing validity evidence to help establish and support a higher or lower degree of validity. In addition, inferences made about potential social consequences of testing can be revisited by findings of new evidence or changes in social conditions. If, in the medical science field, evidence was found that a new surgical procedure was more efficient and safe than the existing procedure, the examination battery should be changed to include a test of the knowledge about this new procedure, rather than continuing to test the old procedure. According to Messick (1992), "validity is an evolving property and validation is a continuing process" (p. 2).

Since validity is an ongoing process, and validity evidence is always incomplete, it is necessary to justify the use of the test and to direct the research needed to gain a better understanding of what the test scores mean and how they can be applied in decision making (Messick, 1992). This should be done by a well grounded validation study that incorporates a myriad of research questions that are based on the balance of evidence available. The purpose of a validation study is to collect various types of evidence to establish to what degree an instrument is functioning as expected, and also to what degree the test scores are being properly used to make inferences. For that reason, a researcher seeks out evidence of different types and from various aspects of the testing instrument and its applied context to make a statement about an instrument's degree of validity. Validity studies help administrators and educators have confidence in the scores and inferences made on the basis of the test.

In the surgeon licensing example, the physician is required to provide more than one source of evidence of high ability to perform surgeries. The state licensing board requires surgeons to pass a written test of thorough medical knowledge. Even though such a test can be said to portray an accurate reflection of the surgeon's passive knowledge in the medical field, patients would be hesitant to trust a surgeon based exclusively on the written test. Instead the governments require performance evidence in the form of a twenty-four month long progressive residence training where the physician is required to demonstrate medical knowledge in real-world situations while being supervised and assessed by an attending physician. Additionally, surgeons have to provide other predictive evidence supporting their

ability as a qualified surgeon. The state licensing board may require a report of good character, and conduct a criminal background check in order to gather information about the surgeon as a responsible and conscientious citizen. These various aspects of the physician's ability provide a more complete indication of his or her potential to meet the standard of a good surgeon. Likewise, a good researcher requires multiple sources of evidence to make a more substantive claim about the validity of an exam.

Messick's Validity Model for Performance Assessment

In the previous section, it was suggested that validity is not necessarily a characteristic of a test, but it rather deals with the interpretation of test scores for a specific purpose. Furthermore, the question of validity incorporates the possible consequences on society caused by implementing decisions made on the basis of test scores. Additionally, the necessity of different sources of evidence in establishing the degree of validity was examined. Another crucial aspect of validity is that it is a unified concept. Traditionally, validity was seen as being composed of three separate types: content validity (the tasks are an adequate representation of the language skills, structures, etc., that the test is supposed to measure), criterion validity (to what extent the test results agree with some other independent indicator of language ability), and construct validity (the test is measuring the desired language ability). However, the test content should be a representation of the construct interpretation and cannot carry on any test purpose on its own. In the same way, criterion-related validity needs to be based on construct-related evidence. For that reason construct validity is a unified concept that comprises both criterion- and content-related evidence to support the meaning and interpretation of test scores. The common view

today is that validity is a unified concept (Messick, 1992; Weir, 2005). The concept of unified validity consists of several types of validity-related evidence. The different types of validity are not three separate types of construct, but rather different complementary components of validity evidence that together establish to what degree the instrument is valid (Messick, 1989a, 1989b). No single source of evidence can establish validity on its own, nor is any one source of validity considered to be superior to another.

Messick's validity model, which includes a two-by-two matrix of the unified validity concept, has been widely recognized by test developers and researchers and is the cornerstone for most validation research (Kunnan, 1998). Messick (1989a) explains that a unified validity framework is constructed

> by distinguishing two interconnected facets of the unitary concept. One facet is the source of justification of the testing, being based on appraisal of either evidence or consequence. The other facet is the function or the outcome of the testing, being either interpretation or use. If the facet for source of justification (that is either an evidential basis or consequential basis) is crossed with the function or outcome of the testing (that is, either test interpretation or test use), we obtain a four-fold classification. (p. 20)

Figure 2.2 shows the unitary validity framework introduced by Messick (1989a). He suggests that the evidential basis for test interpretation proposes that the

evidence be gathered to support the meaning and justification, which means the same as construct validity. Hence, the first cell contains construct validity. The evidential basis of test use comprises, besides construct validity, also the relevance of the scores to the test purpose and the utility of the scores in specified test settings.

	Test Interpretation	Test Use
Evidential basis	Construct Validity (CV)	CV + Relevance/Utility (R/U)
Consequential basis	CV + Value implications (VI)	CV + R/U + VI + Social consequences

Figure 2.2. Facets of Validity (Messick, 1989a, p. 20).

The consequential basis of test interpretation contains both construct validity and value implications of score meaning. Value implications inform not only the construct theory upon which the test is based, but also the meaning of scores based on how the construct was defined. The consequential basis of test use includes construct validity, relevance/utility, value implications and social consequences, which describe the effect the inferences of test scores have on the specified society.

We can see that construct validity appears in all four cells, which suggests that construct validity should be regarded as a superordinate concept embracing all other forms of validity. Evidential basis, consequential basis, and test interpretation and use are interconnected with each other in the process of validation (Messick, 1992).

A Model Describing Language Ability

In the previous sections, it was explained that a precise definition of language ability is essential to making conclusions about an individual's language proficiency

on the basis of performance on a language test. When language ability is defined for the purpose of a specific measurement, a construct is defined. In order to gather the most appropriate evidence to establish to what degree a language test is valid, it is essential to understand what language ability consists of and what a language test is supposed to measure (Palmer, 1991). One of the models of language ability that seems to have attracted the most interest in the past decade is the one proposed by Bachman (1990), shown in Figure 2.3, that was inspired by Canale & Swain (Palmer,1991).

Many language researchers have adapted this model to describe the construct of the different language skills (Fulcher, 2003; Purpura, 2004; Buck, 2001; Alderson, 2000). Since it would be beyond the scope of this research study to discuss the construct of each language skill, I will provide and discuss this fundamental model of language ability. This theoretical model can then function as a guide in defining the construct for any language testing situation.

Bachman (1990) defines language ability as comprising two factors: language knowledge and strategic competence, which is described as a set of metacognitive strategies. According to Bachman and Palmer (1996), "this combination of language knowledge and metacognitive strategies provide language users with the ability, or capacity, to create and interpret discourse, either in responding to tasks on language tests or in non-test language use" (p. 67). They suggest that it is beneficial to be aware of the full range of components of language ability when designing and developing language tests and interpreting the test scores. Even though many of the language tests focus on only one or a few of these areas of language knowledge, the

kinds of test items, tasks, or texts used need to be chosen with an awareness of what other components of language knowledge they may evoke (Bachman and Palmer 1996).

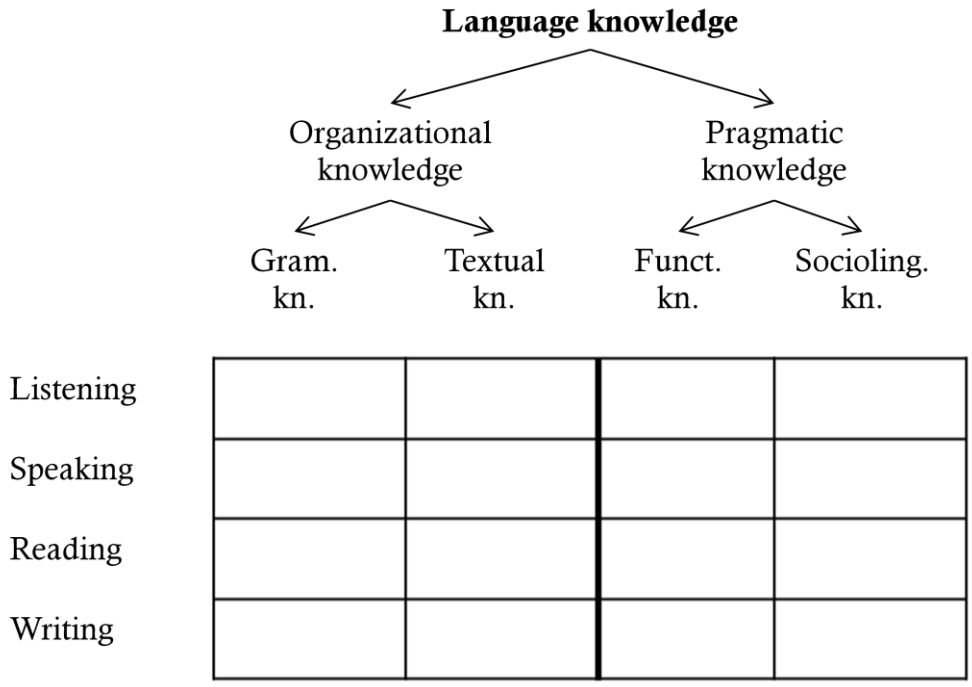


Figure 2.3. Communicative language ability constructs (Palmer, 1991, p. 3).

Language knowledge can be seen as an area of information in memory that that people use through metacognitive strategies to create and interpret communication. Language knowledge consists of two broad components: organizational and pragmatic knowledge.

Organizational knowledge is involved in controlling how utterances or sentences and texts are organized. There are two areas of organizational knowledge: grammatical and textual knowledge. Grammatical knowledge deals with how utterances and sentences are organized and makes use of knowledge about vocabulary, syntax, phonology, and graphology. Textual knowledge deals with the

organization of utterances or sentences to form texts. In order to do that, one retrieves knowledge of cohesion, and rhetorical or conversational organization.

Pragmatic knowledge makes it possible to create or interpret communication by connecting utterances or sentences and texts to their meaning, to the intentions of language users, and to apply the characteristics of language use to a situation. The two areas of pragmatic knowledge are functional and sociolinguistic knowledge. Functional knowledge permits relation of utterances or sentences and texts to discourse goals of language users. Functional knowledge includes four categories of language functions: ideational, manipulative, instrumental, and imaginative. With Sociolinguistic knowledge, a language user is able to create and interpret language that is appropriate for a particular language situation. To be able to do that a language user needs the knowledge of appropriate use of dialects or varieties, registers, natural or idiomatic expressions, cultural references, and figures of speech (Bachman and Palmer, 1996).

In addition to language knowledge, a language user also needs a set of metacognitive components or strategies, "which can be thought of as higher order executive processes that provide a cognitive management function in language use, as well as in other cognitive activities" (Bachman and Palmer, 1996, p. 70). These strategies include being able to decide what one is going to do (goal setting), taking stock of what is needed, what one has to work with, and how well one has done (assessment), and deciding how to use what one has (planning).

Test Score Reliability

The Need for Test Score Reliability

Test validity has been established as the degree to which *reliable* test scores are interpreted correctly and used appropriately for making decisions. In general, reliability deals with the consistency of test scores. Hughes (2003) states that for a test to be valid, it must provide consistently reliable measurements. Nonetheless, a reliable test is not necessarily always valid. This means that a test can provide consistent test scores, which however give the wrong information and might be used for an interpretation that would be inappropriate. In their book *Measurement and Assessment in Teaching,* Linn and Gronlund (2000) use an illustration, as shown in Figure 2.4, that exemplifies the relationship between validity and reliability very well.

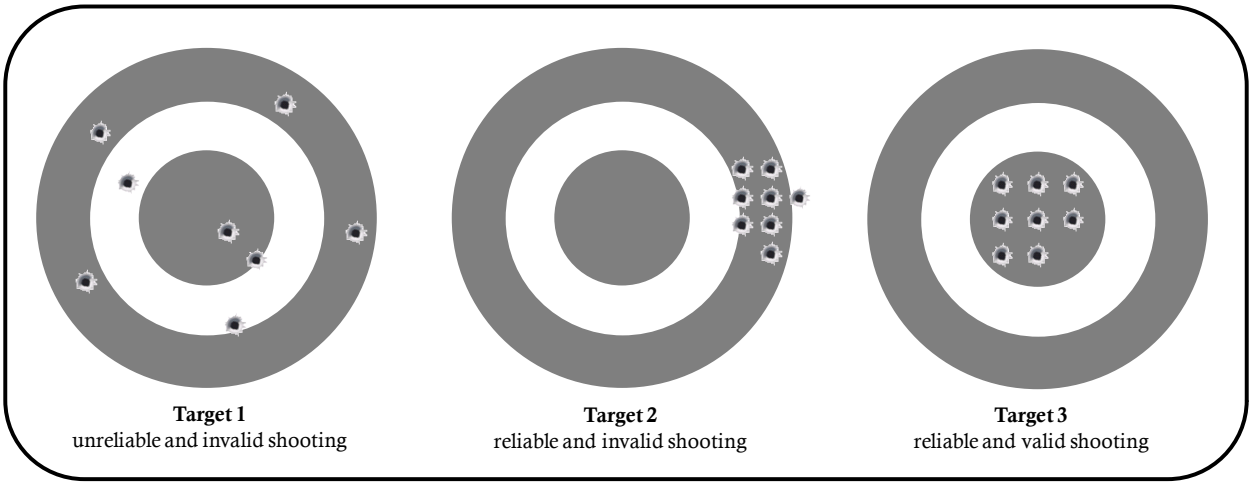


Figure 2.4. Target shooting illustration – relation between reliability and validity (Linn & Gronlund. 2000, p. 75).

Target 1 shows unreliable and invalid shooting; target 2 shows reliable shooting that is not hitting the bulls-eye, meaning that the shooting is not valid. Only

target 3 shows both reliable and valid shooting. This example shows the importance of establishing a degree of validity using reliable test scores to make inferences.

Defining Reliability

After discussing the interconnection of validity and reliability, the exact definition of reliability is easier to understand. Imagine that one hundred students take a 100-item test on spelling rules of English at four o'clock on a Friday afternoon. Since the test is neither extremely difficult nor too easy for these students, they do not all get a zero or a perfect score of 100. Now let us imagine that these students had taken this same test on the previous Tuesday earlier in the day. Even if the test had been a perfect test, administered the same way each time, corrected by the teacher without any bias, and the students had not learned or forgotten anything in the meantime, the students still would not have received the same score on both days. There are many sources that influence test scores to create variance which we cannot prevent. The goal, though, is to construct, administer, and score tests in such a way that students would get a similar score even if the test were administered on different days and at different times (Hughes, 2003). "Reliability thus has to do with the consistency of measures across different times, test forms, raters, and other characteristics of the measurement context" (Bachmann, 1990, p. 24).

Estimating Reliability

It is possible to estimate the degree to which the test scores are reliable by calculating a reliability coefficient. The reliability coefficient of a test can go as high as 1, which would mean that the test would show the exact same test results for a particular set of test takers no matter when the test was administered. A coefficient of

zero would show no reliability. If for example, scores on a test have a reliability coefficient of .89, it would say that the scores are 89% reliable, with 11% measurement error. According to Lado (as cited in Hughes, 2003), vocabulary, structure, and reading tests are considered good, if the reliability coefficient lies between .90 and .99., while listening comprehension tests usually lie in the .80 to .89 range, and oral production tests may range from .70 to .79.

There are three basic strategies for estimating the reliability of most tests: test-retest, equivalent forms, and internal consistency strategies. As the name suggests, with the *test-retest* strategy the same test is given to the same group of students at two different times which are far enough apart time-wise so that the students are not likely to remember the items on the test, but close enough so that the students have not changed in any influential way (like learning more of the language). The two sets of scores are then used to calculate a correlation coefficient. The *equivalent-forms* or sometimes also called parallel-forms reliability is very similar to the test-retest reliability. However, instead of giving the same test twice, two different but uniform tests are given to one group of students. The tester then calculates a correlation coefficient using the two sets of scores.

The most frequently used reliability strategy is the *internal-consistency reliability* coefficient, since it is not necessary to administer tests multiple times. Internal-consistency reliability strategies use only the information internal to a test to estimate the consistency of a test. One of the commonly used methods to calculate internal consistency is Cronbach's alpha coefficient. For detailed instructions on how to

calculate and interpret these coefficients and for more information about the strategies, please refer to Bachmann, 1990; Brown, 2005; Hughes, 2003; Weir, 2005.

Standard Error of Measurement

The reliability coefficients make it possible to compare the reliability of tests, but they do not provide us with concrete information about how close an individual's actual score is to the score he or she might have received on taking the test on a different occasion. The *standard error of measurement* is used to determine a range around a student's actual score within which that student's score would probably fall if he or she were to take the same language test over and over again, without the effect of remembering the items or learning more of the language. Based on the percentages in the normal distribution, we can estimate the probability with which a student's score would fall within one SEM (68% probability), two SEMs (95% probability), or three SEMs (99.7% probability). If, for example, the SEM for a test with a maximum score of 100 is 5, we can conclude the following: a student that scored 75 on the test, would score within a range of one SEM (70 – 80) plus or minus (75+5=80; 75-5=70) 68% of the time if he could take the test over and over again. He would score within the range of 65 – 85, 95% of the time (two SEMs), and in the range of 60 – 90, 99.7% of the time (three SEMs) (Bachmann, 1990; Brown, 2005; Hughes, 2003; Weir, 2005).

Sources of Variance

According to Brown (2005), the performances of students on tests can vary for different reasons. The two general sources that can cause scores to be inconsistent, or in other words, that cause score variance, are (1) variance related to the purpose of

the test, which Brown calls 'meaningful variance', and (2) variance due to other extraneous sources, called 'measurement error' or 'error variance'.

Brown (2005) defines *meaningful variance* "as that variance which is directly attributable to the testing purposes" (p. 170). This basically deals with test validity which has been described and discussed in detail previously. Meaningful variance can be influenced by the test design, the testing method, and the definition of testing purpose, which in other words is the definition of the language construct used as the basis for the test.

Figure 2.5 on the following page lists the potential sources that can generally be associated with *measurement error*, or variance in scores on a test which are not directly related to the purpose of the test.

Rater Reliability

As we can see from the Figure 2.5, error variance can arise for different reasons. These sources of error variance can be grouped into three general areas: variance due to content sampling, which includes test items, variance due to change in conditions in time, which includes examinees, administration and environment, and variance due to individual scorers involved in the scoring process. This source of error variance is also called *rater* or *scorer reliability.* With a test consisting of only multiple-choice items we can assume that scoring of the test would be 'perfect', meaning that if a student performed on the test the same way on every occasion, the student would be given the same score each time. Just as we can estimate the degree to which test scores are reliable by calculating a reliability coefficient, we can estimate the level of agreement given by

the same or different scorers on different occasions by calculating a scorer reliability coefficient.

Variance due to environment	Variance attributable to examinees
• location	• health
• space	• fatigue
• ventilation	• physical characteristics
• noise	• motivation
• lighting	• emotion
• weather	• memory
Variance due to administration procedures	• concentration
• directions	• forgetfulness
• equipment	• impulsiveness
• timing	• carelessness
• mechanics of testing	• testwiseness
Variance due to scoring procedures	• comprehension of directions
• errors in scoring	• guessing
• subjectivity	• task performance speed
• evaluator biases	• chance knowledge of item content
• evaluator idiosyncrasies	
Variance attributable to test and test items	
• test booklet clarity	
• answer sheet format	
• particular sample of items	
• item types	
• number of items	
• item quality	
• test security	

Figure 2.5. Potential sources of error variance (Brown, 2005, p. 172).

The scorer coefficient can be interpreted in a similar way to the test reliability coefficient. In a multiple-choice test as described above, the scorer reliability coefficient would be 1 since it does not involve any judgment from the scorer's side, and could be carried out by a computer. Only subjectivity should cause the scorer

reliability coefficient to drop below 1. For example, for oral performances or tests containing items requiring composition the coefficient will always be under 1. Even though it is not possible to obtain a scorer reliability coefficient of 1 on subjective tests, there are ways to make it sufficiently high for test results to be valuable. There is a very close relationship between scorer reliability and test score reliability. If the scoring of a test is not reliable, the test result cannot be reliable either. The test reliability coefficient will certainly be lower than the scorer reliability coefficient, since there are other sources in addition to scorer (un)reliability that can affect test score reliability.

Ways to Increase Test Reliability

As mentioned above, there are different variables that can cause scores to be unreliable. While test administrators cannot prevent some of the sources, such as personal attributes, from influencing variances, there are many different ways to increase test reliability. In his book, *Testing for Language Teachers,* Hughes (2003) provides a number of practical ways that can help to increase test reliability. In the following section, a few of the strategies that could increase reliability relevant to the German Proficiency Exam are discussed.

1. Take enough samples of behavior: To increase the reliability of a test, it is beneficial to have more items. With respect to the illustration of the target shooting, it would be very hard to determine how reliable a shooter is on the basis of one shot. The same applies to language testing. It is important, though, to keep in mind when adding more items that the added items are independent from each other, meaning that the answer to one item should not build on the information of a previous item.

2. Exclude items which do not discriminate well between weaker and stronger students: A statistical analysis of items can tell whether an item discriminates well or not. The item facility and item discrimination values provide information about individual items.

The *item discrimination* coefficient indicates how well an item discriminates between weak and strong students. The higher the coefficient, the better the item discriminates. The coefficient can range from a maximum discrimination of 1 to a minimum discrimination of zero. If weaker students perform better on an item than stronger students, the coefficient can also show a negative number. The better the items discriminate, the more reliable the scores are. Items with a low discrimination coefficient should be reviewed and improved or taken out of the item pool.

Ebel (as cited in Brown, 1996, p. 70) suggests that the guideline for interpreting item discrimination values should be as follows: values of 0.40 and up are very good items; values from 0.30 to 0.39 are reasonably good but possibly subject to improvement; values from 0.20 to 0.29 are marginal items, usually needing some improvement; and values below 0.19 are poor items, that should either be rejected or revised.

The *facility value*, also referred to as item difficulty, basically indicates what percentage of students got the answer right. If forty-three students out of one hundred answer an item correctly, the item facility value would be .43. That means that 43% of students got the item right. This value provides information on how difficult or easy an item seems to be. The higher the value is, the easier the item is. How the item facility is used depends on the purpose of the test. To develop a proficiency test

that is supposed to identify the top 10% of students, items on the test must be sufficiently difficult. Therefore, the test would need a high proportion of items that have a low facility value. But to develop a test that places students in an array of levels the test should include a wider range of facility values in its items.

A test which has too many easy items usually does not distinguish well between strong and weak examinees because a high percentage of students might answer correctly. Items with either a very high or very low facility value may cause low discrimination values. A test can, in general, discriminate better between weaker and stronger students if the extremely easy and difficult items are left off.

If a test contains multiple-choice items, it is beneficial, in addition to calculating the discrimination coefficient and the item facility value, to conduct an analysis of the distractors. Distractors that are chosen by very few students should be revised, replaced or taken out.

3. Do not allow candidates too much freedom: It is very common in some kinds of language tests to offer students a choice between several questions. In addition to a choice of questions, students are usually allowed a great deal of freedom in the way they answer the question. This, for example is very typical in writing tests. Offering a wide variety of different topics to choose from can have a negative impact on the reliability of the test for different reasons. The questions in themselves can vary in difficulty and require different emphasis in skill in order to perform the task. Topics or items that seem more difficult might be chosen less frequently or topics that seem to deal with an appealing content matter might be too hard for some students to answer. Furthermore, a wide variety of answers can

impose a problem on the scoring procedure. Scoring the compositions all on one topic and allowing a comparison between students as direct as possible will be more reliable. If a choice between questions is offered, the tasks should be worded in a way that it controls more closely what can be written.

4. Write unambiguous items and provide clear and explicit instructions: It is essential to word the items and also the instructions in a clear and explicit manner so that the students understand what exactly the test items are asking for. The items and instructions should be worded in a way so that the students only give answers that are anticipated by the test developer and no answers are given that were not anticipated. For example, in a vocabulary test the test developer should be aware of all the meanings of a word asked for. The item should then be worded so that either all or any of the meanings of the word are acceptable answers. The students should be able to interpret the task and be clear about what they are asked to do. It is better to provide too much information about how to perform the task than too little.

5. Use items that permit scoring to be as objective as possible: Even though multiple-choice items allow for completely objective scoring, it is not necessarily beneficial to use this type of item for all purposes. For some testing situations, multiple-choice items are appropriate to use, but for some other circumstances they are never appropriate. To test writing skill, it would not be suitable to use multiple-choice items that test grammar concepts without having the students actually write a composition. In addition, multiple-choice items are very difficult to write and require extensive pre-testing. If fill-in-the-blank items, open-ended items or essay questions

are chosen, it is important to ensure clear and explicit instructions and guide the responses by not allowing too much freedom.

6. Provide a detailed scoring key based on clearly stated proficiency scales: A fundamental tool that can help with scoring the tests more objectively is to provide a detailed and clear scoring key. If the scoring key is based on a clearly defined proficiency scale or specified levels of proficiency for each language skill, it can provide valuable information and also be a guideline and reference of how much of the skill a student has mastered.

7. Train scorers: Trained scorers are especially important where the scoring is most subjective. The scoring of an essay or an oral performance should only be performed by someone that is very familiar with the proficiency levels and trained on the scoring procedures. After each test has been administered and scored, the patterns of scoring should be analyzed. It is important to see if the rating scale has been applied in a wide range. If, for example, 100 students have been rated on their oral performance and only the top 10% of the rating scale has been used, one should examine whether all of the 100 students really have performed that well on the oral exam, or if the rater has not used the full range of the rating scale to identify the true levels of oral proficiency of the students.

8. Identify candidates by number, not by name and employ multiple, independent scorers: It is unavoidable for a scorer to have expectations for students they know, especially if the teachers are scoring their own students. If there is not a purely objective testing method, it will affect the way they score. To reduce the effect of subjectivity it can be helpful to identify the students by a number and not by name.

Another way to gain more objectivity in scoring, and thus more reliability, is to have at least two trained independent scorers score the oral performance or composition.

To sum up, there are many ways to make tests more reliable. Depending on the circumstances of test development, administration, and scoring it might not be feasible to apply all of these ways to make a test more reliable. It is up to the administrator to consider the circumstances, interpret the scores, and make inferences on the basis of the scores in order to make the decisions on what to put into action to make a test more reliable and thus more valid.

Summary

This chapter has provided information on different approaches to program evaluation and it has described how testing plays a significant role in the process of continuous improvement of a program. Numerous test designs, along with their functions, have been reviewed and the advantages and disadvantages of each have been discussed. An explanation of language ability and a definition of a language construct were given. Benefits of using a clearly defined language construct as a basis of a language test were detailed, including a description of a widely used model of language construct. Following this, the concept of test score validity was defined as a unified construct, as Messick illustrated in his validity model. It was explained that it is crucial to establish the degree of validity by providing different aspects of validity evidence. Finally, the connection between test score validity and reliability was established, giving a definition of reliability and the importance thereof. Different practical strategies were introduced that can support increasing the reliability of test scores.

Current Study

In the vast field of education with its numerous subjects, research always comes at a fast pace, providing the teaching and learning community with new insights, methods, and theories that can be very valuable to the improvement of teaching and learning. It is sometimes a daunting responsibility for those who make decisions, like administrators, teachers, and test and curriculum developers, to continuously keep informed on the developments in the subject area, so that they can provide those that are affected by the decisions with the best education possible.

The aim of this study is to provide current information in the field of language testing and program evaluation that will lead to the improvement of the German Proficiency exam, and thus to the improvement of teaching and learning in the German Section at BYU. For this purpose, this study examines the validity of the GPE scores. As mentioned previously, it is not possible within the timeframe of this research study to conduct a complete validation study, but only to concentrate on specific relevant questions that will help the German Section. Based upon the aim of this study and the review of relevant literature, the following research questions are established to provide guidance for collecting applicable evidence:

I. Does the overall content of the German Proficiency Exam represent general language ability?
II. How reliable is each component of the German Proficiency Exam?
III. In addition the following questions for each of the German Proficiency Exam components are asked:
    1) Listening Comprehension

- How difficult are the items in relation to one another?
- How well do the items discriminate between the different proficiency levels of students?
- Is there sufficient variation in test scores?

2) Reading

- How difficult are the items in relation to one another?
- How well do the items discriminate between the different proficiency levels of students?
- Is there sufficient variation in test scores?

3) Grammar

- How difficult are the items in relation to one another?
- How well do the items discriminate between the different proficiency levels of students?

4) Strong Verbs

- How difficult are the items in relation to one another?
- How well do the items discriminate between the different proficiency levels of students?

5) Vocabulary

- How difficult are the items in relation to one another?
- How well do the items discriminate between the different proficiency levels of students?
- How well do the distractors for each item function?

6) Writing

- How similar are the task options in terms of task difficulty?
- Is there sufficient variation in test scores?

7) Speaking

- Is there sufficient variation in test scores?

Based upon these research questions, a validation study is described in chapter three, which collects the types of evidence that can answer the research questions. The results are presented in chapter four. In chapter five, various types of evidence are discussed and answers to the research questions presented. In addition, chapter five proposes suggestions for improvement of the GPE.

CHAPTER THREE

Research Design

Chapter two provided a theoretical basis for the validation study of the German Proficiency Exam (GPE) for the German Section at Brigham Young University (BYU). Chapter three applies the theory to an investigation into the validity of the GPE. The test, with its components, was given to German students in their senior semester beginning in Fall semester 1998. Qualitative and quantitative data were collected and analyzed for these tests. First, I will describe the participants involved in this study. Then, I will describe each component of the GPE. Finally, I will describe the procedures for the data analysis.

Participants

The participants of the study were 179 adult students at Brigham Young University registered between Fall Semester 1998 and Fall Semester 2007. These students were required to take the German Proficiency Exam during their senior year in order to graduate with a major in German, German Teaching, German Linguistics or German Literature, as well as a minor German Teaching.

Since Fall Semester 1998 the GPE was administered during 17 semesters, of which 10 were fall semesters and 7 were winter semesters. 77 females and 102 males participated in this study. Of the 179 participants, 132 majored in German, 34 in German Teaching, 9 in German Literature, and 1 in German Linguistics. One person had a minor in German, and the majors of two participants could not be determined. Most of the participants were native English speakers. However, there were a few native speakers of other languages, such as German, Swedish, Spanish,

and Russian. Information on other native languages and the exact number of speakers of those languages was not available. In addition, the information about the exact age of the participants was not accessible. Tables 3.1-3.3 present the overall characteristics of the students that have taken the German Proficiency Exam from Fall Semester 1998 to Fall semester 2007.

Table 3.1

Semester Breakdown of Participants (N=179)

Characteristic		n	%
Semester			
	Fall 1998	10	6
	Fall 1999	10	6
	Fall 2000	7	4
	Winter 2001	18	10
	Fall 2001	7	4
	Winter 2002	16	9
	Fall 2002	10	6
	Winter 2003	12	7
	Fall 2003	9	5
	Winter 2004	22	12
	Fall 2004	7	4
	Winter 2005	13	7
	Fall 2005	8	4
	Winter 2006	6	3
	Fall 2006	4	2
	Winter 2007	11	6
	Fall 2007	9	5
	Total	179	100

Table 3.2

Gender Characteristics of Participants (N=179)

Characteristic		n	%
Gender	Female	77	43
	Male	102	57
	Total	179	100

Table 3.3

Breakdown of Participant Degree Emphasis by Gender(N=179)

Characteristic		n	%
Emphasis	German	132	74
	Female	50	28
	Male	82	46
	German Teaching	34	19
	Female	24	13
	Male	10	6
	German Literature[1]	9	5
	Female	2	1
	Male	7	4
	German Linguistic[1]	1	>1
	Female	0	0
	Male	1	>1
	German minor	1	>1
	Female	0	0
	Male	1	>1
	German Teaching minor[1]	0	0
	Female	0	0
	Male	0	0
	other	2	1
	Female	1	>1
	Male	1	>1

Some students were not able to take all components of the GPE, and the data for some students in some components cannot be included in the overall analysis because of missing data parts or incongruence in scoring procedures. For that reason, the number of participants in each component is different. The following Table 3.4 shows the number of participants for each GPE component.

Table 3.4

Number of participants for each component

GPE component	n
Listening	169
Reading	168
Grammar	177
Strong Verbs	179
Vocabulary	179
Writing	152
Speaking	176

Description of the German Proficiency Exam

The purpose of the German program is to facilitate the acquisition and improvement of German language skills (fluency and grammatical knowledge) in all areas of competence--speaking, reading, writing, and listening comprehension--including an understanding of the structure of the German language. All students in the German Section are required to take the same core classes and can choose additional electives depending on their specific major or minor. All students

majoring in in the German program and those with a minor in German Teaching are required to take the German Proficiency Exam (GPE) in order to graduate. This exam is administered in connection with taking the German 400R course. The students are required to take this course during their senior year prior to graduation. The class is meant to prepare the students to take the GPE by making them familiar with the format of the exam and by reviewing the main concepts of the language skills. The content of the GPE, however, is not based on the content or syllabus of German 400R.

The purpose of the GPE is to assess the level of German proficiency of all students in the German program at the time of graduation. The GPE is meant to inform students of their exam performance in relation to other students in their testing group. Further, it is supposed to provide students with some concrete indication of their German proficiency, which they may show to employers or other interested parties. Finally, the GPE functions as a basis for assessing the proficiency levels of students graduating from the German program. Generally, it provides information on how fluent the students are and how well they perform in the different language areas of speaking, listening, reading, writing, vocabulary, and grammar.

For that reason, there are seven components to the GPE: listening comprehension, reading, grammar, strong verbs, vocabulary, writing, and speaking. In order to answer the research questions, each instrument component, with its procedures and data analysis, will be described in the following section.

Listening Comprehension

The purpose of the listening comprehension component is to determine how well the students understand what they hear in German. The component is a paper-based test and is conducted separately from the other six components. The test consists of eleven short essay question items. Each item is a question pertaining to the audio passage the students listen to. First, the students are given the paper test with the eleven item questions. They get a few minutes to read through the questions, so they know what they need to look for. Then the CD selection is played once, and notes can be taken. Students are given time to answer the questions by writing in the space underneath each question. At the end, the audio passage is played again and the students can make final revisions to their answers.

The tests are collected by the teaching assistant administering the test and given to a grader. Using the scoring key provided, the grader scores each question. The professor in charge of the German Proficiency Exam then reviews the initial corrections and gives a final score. The maximum score that can be given for each answer is 10 points. The maximum total score of the listening comprehension component is 110 points. In order to facilitate data analysis, the scores of this GPE component were transferred to an Excel file.

Reading

The purpose of the reading component of the GPE is to determine how well students understand what they read in German. The component is a paper-based test and is conducted separately from the other six components. The test component consists of one longer segment of a newspaper article that is divided in seven separate

parts. The students read each part of the article and are asked to write a short paragraph about the key points of each article part in order to show how much they have understood of what they have read.

The reading component is corrected by a professor who checks each task paragraph for completeness of content and gives a final score. The maximum score that can be given for each task is 10 points. The maximum total score of the reading component is 70 points. In order to facilitate data analysis, scores of this GPE component were transferred to an Excel file.

Grammar

The purpose of the grammar component of the GPE is to determine how well students analyze sentences and understand which structural forms are required in a given sentence. The component is a paper-based test and is administered separately from the other six components. The grammar test consists of a text with blanks for 54 different sentence-structure parts. Each blank is identified by a number from one through fifty-four. For the blanks, an answer column is provided, consisting of fifty-four numbered blank spaces. The students are advised to first read the entire passage. They then go back and write the appropriate missing word, words, or word parts into each corresponding blank space.

The grammar component is corrected by a grader who checks the answer for each blank test item. Each item can either be correct or incorrect. If meaning, form, capitalization, and spelling are correct, one point is given. Otherwise, no point is given. The maximum total score of the grammar component is 54 points. In order to

facilitate data analysis, scores of this GPE component were transferred to an Excel file and converted to binary data.

Strong Verbs

The purpose of the strong verbs component of the GPE is to determine how well students know the principle parts of strong verbs. The component is a paper-based test and is administered separately from the other six components. The students are given a table illustrated on a paper consisting of 10 columns and 9 rows. For the purpose of this study, each of the columns is identified by a capitalized letter and the rows are identified by a number, as shown in Figure 3.1.

	A	B	C	D	E	F	G	H	I	J
	Infinitive	1st present	2nd present	3rd present	du imperative	preterite	subj. II	participle	aux	English
1	A1	B1	C1	D1	E1	F1	G1	H1	I1	J1
2	A2	B2	C2	D2	E2	F2	G2	H2	I2	J2
3	A3	B3	C3	D3	E3	F3	G3	H3	I3	J3
4	A4	B4	C4	D4	E4	F4	G4	H4	I4	J4
5	A5	B5	C5	D5	E5	F5	G5	H5	I5	J5
6	A6	B6	C6	D6	E6	F6	G6	H6	I6	J6
7	A7	B7	C7	D7	E7	F7	G7	H7	I7	J7
8	A8	B8	C8	D8	E8	F8	G8	H8	I8	J8
9	A9	B9	C9	D9	E9	F9	G9	H9	I9	J9

Figure 3.1. Strong verbs item overview.

Each of the numbered rows belongs to one strong verb. In the cells of column A the students need to fill in the infinitive of the strong verb. For column B, C, and D the strong verb form in the first person, second person and third person of the present tense respectively needs to be filled in. In column E, the second person singular imperative form of the verb is required, and the subjunctive II form is required in column G. Columns H and I need to be filled in with the correct past participle form and corresponding auxiliary verb. In the last column (J) the students need to provide

the English meaning of the German strong verb. In the cells which are shaded in Table 2 the specific form of the strong verb is already provided. The students use those clues to help them figure out the rest of the forms asked for in the remaining cells.

Each cell of the strong verb table is checked by a grader for accuracy. Each item can either be correct or incorrect. If the cell contains the correct form, one point is given. If the cell does not contain the exact form, no point is given. The maximum total score of the strong verbs component is 81 points. In order to facilitate data analysis, scores of this GPE component were transferred to an Excel file and converted to binary data. Each item is identified by a capital letter and a number.

Vocabulary

The vocabulary component of the GPE is used to estimate the vocabulary size of each student. How exactly this is done will be explained at the end of this chapter. The vocabulary component consists of 100 multiple-choice items with four distractors each. The students are given a test paper with the multiple-choice items and a bubble sheet, on which they mark the correct distractor options. For each item one German word is given. The four distractors a), b), c) or d) are each an English word of which one, two, three or all can be the correct meaning of the German word. For every item the students mark all the distractors they think contain the meaning of the German word.

The vocabulary component is corrected by a Scantron machine that reads all the bubble sheets. An item is marked 'correct' and a point is given only if all the corresponding meanings are chosen by the student. Otherwise no point is given. If a

student, for example, has marked two of three possible meanings the item is marked incorrect, and no point is given. The maximum total score of the vocabulary component is 100 points. In order to facilitate data analysis, scores of this GPE component were transferred to an Excel file and converted to binary data.

Writing

The purpose of the writing component of the GPE is to determine how fluently and correctly the students are able to write a composition. The component is a paper-based essay question test and is conducted separately from the other six components. For the writing test, the students are given three topics to choose from and are expected to write an essay of 200-250 words on one. For the first topic choice, a statistical graph is given that illustrates an issue concerning the German society. The students are asked to analyze the information given in the graph and to give their opinion about the issue addressed. The second topic provides two quotes for which the students write their personal opinion. In addition to that, they answer a few more questions about the quotes. For the third topic option, the students can choose to write a letter to the editor in reaction to an article that is provided. The format of the letter is given, and a few questions are provided for consideration when writing the letter.

The grading procedure is conducted by a professor, who checks the essays for correctness in the following areas: word endings, word order, verb forms, idiomatic phrases, spelling and punctuation, and content. The maximum score given for the content is 25 points. For each of the other scoring areas 10 points can be given. The

total score possible for the writing component is 75 points. In order to facilitate data analysis, scores of this GPE component were transferred to an Excel file.

Speaking

The purpose of the oral exam component is to determine how fluently and correctly the students speak German. The oral component is conducted as an interview separately from the other six components. The students sign up for the oral exam time two weeks in advance. The duration of the oral exam is approximately 20 minutes. The examinees are required to come to the exam administration location 20 minutes prior to their scheduled exam time. During these twenty minutes they choose one picture from a file of pictures and receive a sheet of paper with five exam questions, from which they choose one. They are allowed to prepare for the actual exam in any way they find appropriate (e.g. using dictionary). The oral exam is divided into two parts: For the first part, the students describe and discuss the picture they have chosen prior to the exam. During the second part, they respond to the question they have chosen beforehand and discuss it with the testers.

For each oral exam two professors from the German Section score the performance of the examinee. One professor tracks the grammar usage, while the other tracks pronunciation. For the tracking of grammar usage and pronunciation, the testers use the 'Speaking: Grammar Usage' diagnostic instrument sheet and the 'Speaking: Pronunciation' diagnostic instrument sheet that are depicted in Appendix A and B. Each student is evaluated using a 20 point scale which is based on the Expanded Description of Proficiency Levels described in Appendix C. There are four

proficiency levels ranging from high to low. The four score ranges are 0-5 points, 6-10 points, 11-15 points, and 16-20 points. However, for the purpose of the GPE oral component this proficiency scale was modified to fit the needs of the exam. Since the general proficiency level of graduating German students usually does not include very low proficiency, the rating scale from 12 to 20 points is used. The modified rating scale is depicted in the Guidelines for Evaluating Proficiency Exam Orals in Appendix D.

The maximum score that can be obtained for grammar usage, as well as pronunciation is 20 points each, which adds up to a total speaking component score of 40 points possible. In order to facilitate data analysis, scores of this GPE component were transferred to an Excel file.

Summary Score of the German Proficiency Exam

In Table 3.5 on the next page an overview of each component of the German proficiency exam is illustrated. A summary of scores obtained in the German proficiency exam is given in form of a Summary Score Sheet shown in Appendix E. First, the scores for each exam component are transformed into a percentage score and marked in the corresponding column. The percentage score the students receive for their vocabulary component correlates with an estimated number of thousands of words the student knows. For example, if a student scores with a percentage of 80%, it is estimated that she knows about 16,000 words. A separate percentage score for grammar usage and pronunciation is provided. Then, the overall average for the corresponding student is calculated using the percentage scores of all GPE components except of the vocabulary component. The vocabulary component is not

included in the calculation, since the scores the students get in the vocabulary component are usually very low and lower the overall percentage score too much.

Table 3.5

GPE component overview

GPE component	Number of items	Points per item	Total score possible
Listening	11	10	110
Reading	7	10	70
Grammar	54	1	54
Strong Verbs	81	1	81
Vocabulary	100 400 distractors	1	100
Writing	3 topic options 6 scoring areas: word endings word order verb forms idiomatic phrases spelling/punctuation content	 10 10 10 10 10 25	75
Speaking	grammar usage pronunciation	20 20	40
Total Score of the German Proficiency Exam			**530**

The overall average of the student is written in the far right column. The four cells to the right of the overall average of the student give information about an estimated level the student would obtain for the ZDaf (Zertifikat Deutsch als Fremdsprache – certificate for German as foreign language), for the ZMP (Zentrale Mittelstufenprüfung – proficiency exam in German taken by German students at the

end of middle-school), of the ZOP (Zentrale Oberstufenprüfung – proficiency exam in German taken by German students at the end of their senior year), and of the OPI (Oral Proficiency Interview). In the top right cell, each student's performance is shown relative to the whole group of students that took the test that specific semester. Lastly, the average total score for the student and the average total score for the whole group is provided.

Test Analysis

In order to determine how well the GPE represents a general language construct, all test content, including all the test components, were compared to the major language ability construct categories derived from the model for language ability by Bachman (1990). First, the components of the GPE were listed. Then, the language ability construct areas were adapted from Bachman's (1990) model of language ability reviewed in Chapter Two, which can function as the basis for defining the language construct of a test. Under each of these language ability areas -- which are grammatical knowledge, textual knowledge, functional knowledge, and sociolinguistic knowledge -- their corresponding categories were listed. Finally, the content of each GPE component was analyzed and the categories covered by the content of each component were checked. The percentage of the language ability categories covered by the content of the GPE components was calculated.

In order to estimate the reliability of the test scores, Cronbach's Alpha was calculated for the listening comprehension exam, the reading exam, the grammar exam, the strong verbs exam, and the vocabulary exam. In addition, for each of these exams, overall descriptive statistics including mean and standard deviation of

total scores were examined. Then, the item facility and item discrimination for each item on the tests were calculated. A distractor analysis was also completed for the vocabulary exam.

For the writing exam, descriptive statistics of the total scores for each essay topic option were calculated. This included the scores of each scoring area: word endings, word order, verb forms, idiomatic phrases, spelling and punctuation, and content. In order to determine whether or not the topic options were of equal difficulty, an ANOVA on the three topic option total scores was completed. If the ANOVA showed a difference in difficulty a post-hoc Tukey's pairwise comparison will be conducted.

For the speaking exam, descriptive statistics of the total scores for the grammar usage and pronunciation were calculated.

CHAPTER FOUR

Results

In this chapter the results of the test analysis are presented in order to answer the research questions concerning the validity of the German Proficiency Exam (GPE). First, the content of each test component is analyzed. Then, Cronbach's alpha reliability coefficient of the listening, reading, grammar, strong verbs, and vocabulary components is presented. Finally, the results of the data analysis for each component of the GPE are presented.

Content Analysis

The purpose of the content analysis was to examine how fully all test content, including all the test components, covers the major language ability construct categories derived from the model for language ability by Bachman (1990).

Figure 4.1 presents an overview of the general construct areas covered by the GPE. The construct areas are listed in the rows and the GPE components are listed in the columns. The construct areas covered by a GPE component are marked with a check mark. We can see that all the general construct areas are covered by at least one of the GPE components.

Test Score Reliability

In order to estimate the reliability of the listening comprehension exam, the reading exam, the grammar exam, the strong verbs exam, and the vocabulary exam of the GPE, Cronbach's alpha reliability coefficient was calculated.

	Listening	Reading	Writing	Speaking	Gram	Strong Verbs	Vocab.
Grammatical knowledge							
Vocabulary	✓	✓	✓	✓		✓	✓
Syntax			✓	✓	✓		
Phonology/ graphology	✓		✓	✓			
Textual knowledge							
Cohesion		✓	✓	✓			
Rhetorical or conversational organization			✓	✓			
Functional knowledge							
Ideational functions			✓	✓			
Manipulative functions			✓	✓			
Heuristic functions			✓	✓			
Imaginative functions			✓	✓			
Sociolinguistic functions							
Registers	✓			✓			
Natural or idiomatic expressions		✓	✓				
Cultural references/figures of speech	✓	✓		✓			

Figure 4.1. Overview of language construct area coverage.

As mentioned in Chapter 2, according to Lado (as cited in Hughes, 2003), vocabulary, structure, and reading tests are considered good if the reliability coefficient lies between .90 and .99. Listening comprehension tests usually lie in the .80 to .89 range, and oral production tests may range from .70 to .79. Table 4.1 shows that the strong verbs component and the vocabulary component follow this trend; the other three components, however, do not. The reliability coefficient of the grammar, reading, and listening components are lower than expected. Usually, the fewer items a test has, the lower the reliability coefficient is. Further, it is common to have lower reliability estimates with restricted ability range of examinees when correlation statistics are used to estimate reliability. Since the GPE is focused towards students prior to graduation, we can assume that the general ability level is slightly higher. However, it is necessary to take a closer look at the descriptive statistics and the item analysis in order to determine any specific source of variance.

Table 4.1

Cronbach's alpha

	Listening	Reading	Grammar	Strong Verbs	Vocabulary
Cronbach's alpha	0.75	0.89	0.88	0.92	0.94

Listening

Table 4.2 presents descriptive statistics for the listening comprehension component. The descriptive statistics in this table include the mean, standard deviation, the minimum and maximum scores, and the standard error of measurement.

We can see that the mean of the listening component is very high and the standard deviation suggests that the dispersion around the mean is not very wide, even though the range of scores is 90 points.

Table 4.2

Descriptive statistics of the listening component

	Listening
Mean	103.4 (94%)
Standard deviation	11.83
Minimum	20
Maximum	110
SEM	5.84

The item analysis provides more information. When we look at the item analysis in Table 4.3, it shows the reason for the high overall mean and the low standard deviation. The mean of each item is around 9 points out of 10 total points possible for each item. Even though the range is 10 points the scores of each item concentrates around the median of 10. The high item facility suggests that the item is very easy and most students get a score of 10 points on each item. The low item discrimination shows that the items do not discriminate well between weaker and stronger students. Reasons for the high item facility, mean, and median, and low item discrimination may be that the items are too easy, there is a lot of variance due to the scoring procedures, or the scoring procedures are not clearly defined.

Table 4.3

Item analysis of the listening component items

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11
Mean	9.14	9.73	9.02	9.06	9.86	8.86	9.76	9.47	9.77	9.73	9
Mode	10	10	10	10	10	10	10	10	10	10	10
Median	10	10	10	10	10	10	10	10	10	10	10
Minim.	0	0	0	0	0	0	0	0	0	0	0
Max.	10	10	10	10	10	10	10	10	10	10	10
IF	0.91	0.97	0.90	0.91	0.99	0.89	0.98	0.95	0.95	0.97	0.90
ID	0.26	0.08	0.26	0.27	0.04	0.26	0.07	0.13	0.05	0.07	0.27

Reading

Cronbach's alpha for the reading test (.89) is just under the recommended level. The mean shown in Table 4.4 is relatively low with 68.3% due to some harder items (3,4,5, and 6), as we can see in Table 4.5 under item facility.

Table 4.4

Descriptive statistics of the reading component

	Reading
Mean	47.81 (68.3%)
Standard deviation	10.71
Minimum	8
Maximum	70
SEM	3.59

However, item 7 discriminates well between stronger and weaker students, and items 3, 4, and 6 are moderately good. Items 1, 2, and 5 don't discriminate as well between students as the other items and should be considered for revision.

Table 4.5

Item analysis of the reading items

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Mean	8.71	7.59	5.74	5.8	6.71	6.2	7.07
Mode	10	8	6	6	7	7	10
Median	9	8	6	6	7	7	7
Minim.	2	1	0	1	0	0	0
Max.	10	10	10	10	10	10	10
IF	0.87	0.76	0.57	0.58	0.67	0.62	0.71
ID	0.22	0.23	0.37	0.35	0.25	0.37	0.40

Grammar

The descriptive statistics in Table 4.6 show a relatively good standard deviation and a good standard error of measurement. Nevertheless, the slightly lower Cronbach's alpha value and the moderately low mean point towards problematic items, and thus a source of lower reliability.

Items that seem easy with a very low discrimination value need to be revised or taken out of the item pool. The items marked in gray in Tables 4.7-4.8 on the next pages fall under that category. Very difficult items with an item facility value below 0.40 need to be examined also.

Table 4.6

Descriptive statistics of the grammar component

	Grammar
Mean	39.79 (73.7%)
Standard deviation	7.35
Median	41
Mode	45
Minimum	16
Maximum	53
SEM	2.59

Items that are difficult but have high item discrimination may be useful to keep on the test. Yet, items which are both difficult and have low discrimination values, such as items 31, 34, 41, and 54 are difficult for both strong and weak students and need to be reviewed.

Table 4.7

Item facility and discrimination values of the grammar component

Item	1	2	3	4	5	6
Item facility	0.98	0.75	0.86	0.65	0.87	0.86
Item discrimination	0.05	0.49	0.32	0.34	0.29	0.14
Item	7	8	9	10	11	12
Item facility	0.45	0.68	0.91	0.92	0.69	0.95
Item discrimination	0.58	0.27	0.20	0.17	0.39	0.08
Item	13	14	15	16	17	18
Item facility	0.95	0.97	0.95	0.96	0.85	0.98
Item discrimination	0.12	0.10	0.14	0.12	0.03	0.05

Table 4.8

Item facility and discrimination values of the grammar component, continued

Item	19	20	21	22	23	24
Item facility	0.97	0.51	0.63	0.84	0.95	0.65
Item discrimination	0.07	0.61	0.54	-0.02	0.51	0.51
Item	25	26	27	28	29	30
Item facility	0.64	0.83	0.72	0.93	0.91	0.87
Item discrimination	0.41	0.34	0.41	0.20	0.25	0.29
Item	31	32	33	34	35	36
Item facility	0.19	0.49	0.95	0.11	0.64	0.92
Item discrimination	0.24	0.73	0.12	0.17	0.71	0.19
Item	37	38	39	40	41	42
Item facility	0.62	0.69	0.98	0.98	0.25	0.68
Item discrimination	0.73	0.56	0.05	0.05	0.08	0.63
Item	43	44	45	46	47	48
Item facility	0.84	0.82	0.63	0.34	0.49	0.42
Item discrimination	0.34	0.37	0.10	0.20	0.20	0.71
Item	49	50	51	52	53	54
Item facility	0.66	0.92	0.87	0.80	0.69	0.10
Item discrimination	0.64	0.24	0.25	0.41	0.34	0.12

Strong Verbs

The descriptive statistics of the strong verbs component are presented in Table 4.9. The mean score is slightly low, but the other values should not be of concern. The coefficient alpha for the strong verbs component was very good, with a value of 0.92. However, the item analysis provided in Table 4.10 on page 69 shows that a very good coefficient alpha and relatively good descriptive statistics values can be misleading. Each numbered row asks for different grammatical forms of one verb,

and the columns represent the different grammatical forms of the verb asked for. The cells with a dash '-', are those cells where the answer is already provided to give the students a clue about the verb asked for.

Table 4.9

Descriptive statistics of the strong verbs component

	Strong Verbs
Mean	63.74 (78.7%)
Standard deviation	10.84
Median	65
Mode	68
Minimum	32
Maximum	81
SEM	3.15

When evaluating the reliability of a test, it is crucial to not only look at descriptive statistics and reliability coefficients, but to also do an item analysis and consider the content and rating procedures. All of these types of evidence together give an estimate of the validity of test scores.

The item facility and item discrimination of the strong verbs suggest that all the grammatical forms of verb number 7 and verb number 9 discriminate well and have a good item facility value. These values, however, are misleading if we consider the scoring procedure. Each form of a verb (items with the same number ranging from A-J) is marked as correct only if the answer is exact, meaning the correct form of the correct German verb. For example, a lot of students could not guess the correct German verb from the clue given for verbs number 7 and number 9, but most

Table 4.10

Strong Verbs Item Analysis

		A Infinitive	B 1st present	C 2nd present	D 3rd present	E *du* imperative	F preterite	G subj. II	H participle	I aux	J English
1	Item facility	-	0.98	0.49	0.66	0.95	0.80	0.67	0.99	0.82	0.98
	Item discr.	-	0.05	0.47	0.63	0.10	0.41	0.63	0.00	0.07	0.07
2	Item facility	0.99	-	0.42	0.40	0.95	0.79	0.63	0.96	0.97	0.96
	Item discr.	0.03	-	0.61	0.58	0.03	0.44	0.73	0.05	0.02	0.12
3	Item facility	0.97	0.96	-	0.96	0.96	0.96	0.47	0.78	0.81	0.99
	Item discr.	0.00	0.03	-	0.05	0.12	0.10	0.49	0.07	0.22	0.00
4	Item facility	0.90	0.84	0.77	-	0.85	0.93	0.80	0.90	0.65	0.99
	Item discr.	0.15	0.12	0.14	-	0.15	0.14	0.39	0.22	0.05	0.00
5	Item facility	0.94	0.95	0.96	0.96	-	0.88	0.72	0.74	0.98	0.96
	Item discr.	0.12	0.10	0.08	0.12	-	0.22	0.51	0.32	0.03	0.00
6	Item facility	0.92	0.92	0.88	0.91	0.91	-	0.83	0.89	0.90	0.90
	Item discr.	0.22	0.17	0.25	0.24	0.22	-	0.39	0.20	0.20	0.25
7	Item facility	0.53	0.51	0.32	0.39	0.40	0.61	-	0.46	0.35	0.43
	Item discr.	0.69	0.75	0.66	0.71	0.64	0.68	-	0.73	0.63	0.73
8	Item facility	0.97	0.93	0.80	0.80	0.89	0.61	0.47	-	0.94	0.97
	Item discr.	0.05	0.07	0.37	0.44	0.08	0.37	0.56	-	0.02	0.08
9	Item facility	0.71	0.68	0.66	0.69	0.64	0.69	0.60	0.69	0.93	-
	Item discr.	0.51	0.54	0.61	0.56	0.58	0.56	0.71	0.49	0.20	-

of them wrote the correct grammatical form of the wrong verb. Yet, no credit was given, even though the students had the knowledge of how to build the grammatical form correctly. If partial credit were to be given for knowing the grammatical form, the item discrimination, the item facility, and therefore the values of the descriptive statistics would be very different.

Most of the items of the grammatical forms asked for in columns A, B, E, H, I, and J show very low discrimination (except items of verb 7 and 9), and are thus not very useful. The columns C, D, F, and G contain more items that discriminate better, which points out that the grammatical forms asked for in these columns are more challenging. Nonetheless, due to the scoring procedure, all these values need to be interpreted with caution.

Vocabulary

The descriptive statistics presented in Table 4.11 show a very low mean, with an also low median of 45. The maximum score of 86 and the minimum score of 15 point towards a very difficult test component. The high value of the Cronbach's alpha coefficient, however, is due to the large number of items.

Appendix F provides us with the item analysis. The column with the heading 'item' contains the item number. The column with the heading 'IF' contains the item facility, and the following column informs us about the item discrimination. The column specified with 'Distractor' lists distractors A through D for each item. The last column contains the distractor value '*p*', which shows how many of all the students that took the vocabulary component test chose each distractor. If, for example, the value p of a distractor is 0.34, 34% of the students marked that specific

distractor. The values highlighted in gray are the distractors that contain the correct meaning.

Table 4.11

Descriptive statistics of the vocabulary component

	Vocabulary
Mean	48.42 (48.4%)
Standard deviation	16.22
Median	45
Mode	53
Minimum	15
Maximum	86
SEM	4.02

As we can see in Appendix F, the discrimination value of 28 items lies below 0.19 and should either be rejected or revised. The discrimination ability of 11 items is marginal with a value between 0.20 and 0.29 and are in need of some improvement. 15 items discriminate reasonably well with a value between 0.30 and 0.39, but could possibly be improved. There are 46 items remaining, which seem to be very good items. Table 4.12 provides the information which items fall in these four categories.

In the previous section, it was mentioned that it is important to also look at the item analysis and consider the content and rating procedures to estimate to what degree the test scores are valid. As with the strong verbs component, the scoring process of the vocabulary component has a negative effect on the reliability of test scores. Each item has four distractors of which one, two, three or all four can contain the correct meaning of the German word. The students mark each distractor that

they think contains the correct meaning. An item is marked 'correct' and a point is given only if all the corresponding meanings are chosen by the student.

Table 4.12

Vocabulary item discrimination value categories

Item discrimination value range	No. of items in range	Item number
< 0.19	28	1-5, 8, 14, 22, 24, 29, 34, 63, 65, 81, 85, 87-88, 90-100
0.20<>0.29	11	6, 9-10, 15, 19, 25, 36, 58, 72, 78, 82
0.30<>0.39	15	7, 18, 21, 40, 54-55, 64, 68-69, 76-77, 83-84, 86, 89
>0.40	46	11-13, 16-17, 20, 23, 26-28, 30-33, 35, 37-39, 41-53, 56-57, 59-62, 66-67, 70-71, 73-75, 79-80

Otherwise no point is given. If a student, for example, has marked two of three possible meanings, the item is marked incorrect and no point is given. Because no credit is given for knowing at least one of multiple possible meanings of the German word, the total scores of the vocabulary component does not show how much an individual student really knows. In addition, if a student chooses two or more distractors, of which one is the correct distractor, for items that only have one correct distractor, no point is given. For that reason, the values of item facility and item discrimination of especially the items with multiple correct distractors, and items containing only one correct distractor where students have marked multiple distractors, are not a true representation of the difficulty and the discrimination ability of the item. The items that have multiple correct distractors are: 2, 3, 4, 5, 14, 16, 17, 58, 60, 66, 68, 71, 73, 76, 78, 79, 87, 94, and 96. Particular caution should be given to the item facility and item discrimination value of these items.

In addition, the item facility and item discrimination values of multiple-choice items are mainly influenced by the distractors. Appendix F provides us the information about the percentage of responses for each option of all 400 distractors. Bachman (2004) suggests that every distractor should attract some responses, or it is not doing its job as a distractor. Each distractor should have a percentage value of at least 0.10. If a distractor has a value below 0.10 it should be revised. Almost all items of the vocabulary component contain at least one distractor with a value below 0.10. Making distractors more plausible might help to increase discrimination values of items. Items that contain distractors that attract almost as many or more answers than the correct distractors should also be revised. That is the case with 22 items of the vocabulary component.

Writing

The following Table 4.13 provides the descriptive statistics for each of the three topics of the writing component.

Table 4.13

Descriptive statistics of total scores of the three writing component topics

	Topic 1	Topic 2	Topic 3
Mean	66.04 (88.1%)	68.82 (91.8%)	67.91 (90.5%)
Median	66	69	69
Mode	66	72	70
Standard deviation	4.04	3.40	3.92
Range	19	15	20
Minimum	54	60	55
Maximum	73	75	75
Count	55	62	34

For all three topics the mean is relatively high, with the median score being 66 and 69. In addition we can observe a very narrow range of scores, which contributes to a narrow spread, indicated by the standard deviation values. We can see that the higher the mean the lower the standard deviation.

If we look at the descriptive statistics of the separate scoring areas in Table 4.14, the reason for a high mean and low range becomes clear. The mean of each of the scoring areas makes up the high overall mean. The mode of the scoring areas, excluding the content area, is either 8, 9 or 10, and the range doesn't exceed 5, which indicates that the scoring range from 1-10, or 1-25 for the content area, is not sufficiently used.

Table 4.14

Descriptive statistics of writing component scoring areas

	Endings	Word order	Verb forms	Idioms	Spelling	Content
Mean	8.08 (80.8%)	9.39 (93.9%)	9.64 (96.4%)	7.89 (78.9%)	8.14 (81.4%)	24.45 (97.8%)
Median	8	10	10	8	8	25
Mode	8	10	10	8	9	25
Range	4	3	3	5	4	5
Minimum	6	7	7	4	6	20
Maximum	10	10	10	9	10	25

In order to determine whether or not the topic options of the writing component were of equal difficulty, an ANOVA on the three topic option total scores was completed. The results of the ANOVA are shown in Table 4.15. The

columns indicate the sources of variation, whereas the rows contain the values between or within groups or the total.

Table 4.15

ANOVA of the three topics of the writing component

Source of variation	Variability (SS)	Error terms (df)	Variance (MS)	F-ratio	p-value
Between groups	230.45	2	115.22	8.14	0.00
Within groups	2095.71	148	14.16		
Total	2326.16	150			

The F-ratio is calculated by dividing the "between goup" variance over "within group" variance. If there is no difference between topics the F-ratio value would be 1. Large differences between groups, or in this case between the three topic groups, produce a large F-ratio. We can see that there is a large difference between groups, since the F-ratio value is 8.14.

The results of the ANOVA indicate that at least one pair of the topics has significantly different means. A post-hoc analysis using Tukey's pairwise comparisons reveals that the mean for Topic 1 was significantly different form Topic 3. Figure 4.2 demonstrates the overlap of the three topics. There is much overlap between topic 2 and 3 and a little overlap between topics 1 and 3. Topic 1 and 2, however, do not overlap at all. Therefore topic 1 is most different of all three topics.

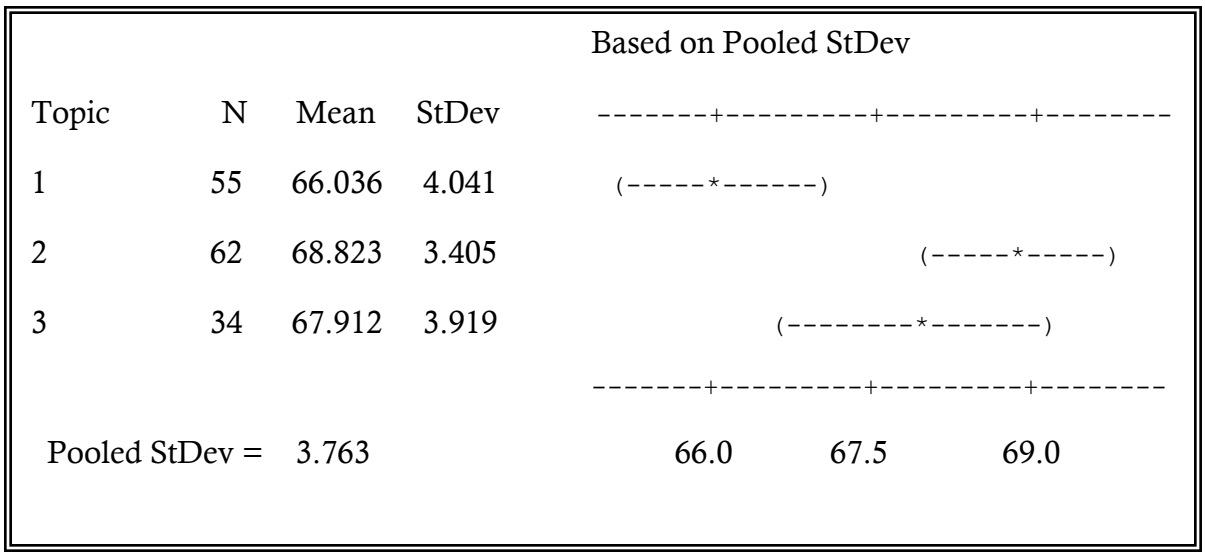


Figure 4.2. Comparison of writing component topics.

Speaking

The following Table 4.16 provides the descriptive statistics for the total scores of the speaking component. The total score consists of the average of the score each student receives for grammar and pronunciation. The mean is moderately high, but no source of concern. The mode shows that most students receive a score of 19, which is very high for a maximum score of 20.

Table 4.16

Descriptive statistics of the speaking component

	Speaking
Mean	17.29 (86.5%)
Median	18
Mode	19
Standard deviation	1.92
Range	9
Minimum	11
Maximum	20

The range of 11 also points out that the rating scale is not being fully used. Even though we would expect more scores in the upper range for students of higher level of ability, there is a concentration around almost the maximum score.

The graph presented in Figure 4.3 shows the frequency of scores for both grammar and pronunciation.

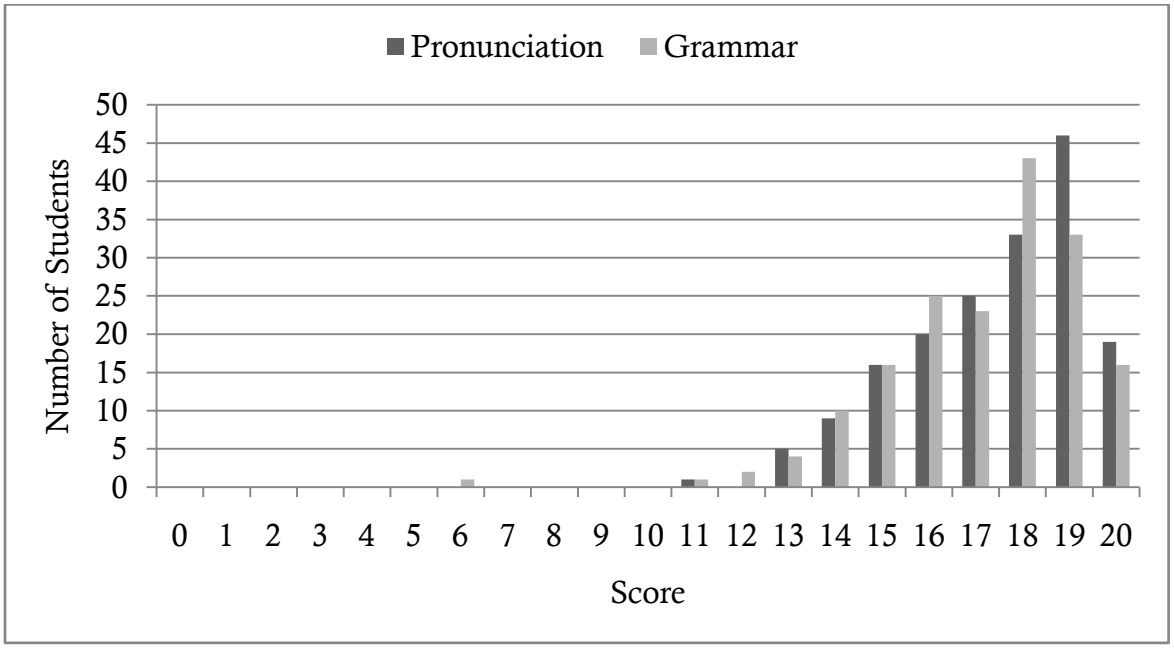


Figure 4.3. Speaking pronunciation and grammar score distribution.

The graph shows skewness to the far right for the grammar and the pronunciation scores. There are more high scores for pronunciation than grammar, indicating that the students are either better at pronunciation than grammar or the raters score pronunciation more leniently. In addition, we can see that there is not a wide spread of scores, with only one outlier at score 6.

Summary

This chapter has presented the results of the data analyses that collected the relevant evidence to answer the research questions in Chapter five. Chapter five will

provide a discussion of these findings with suggestions for improving the GPE and limitations of this study for the direction of future research.

CHAPTER FIVE

Discussion and Conclusion

Evidence has been collected for the validation study of the German Proficiency Exam. The purpose of this chapter is to answer the research questions on the basis of that evidence, and then to discuss implications, limitations, and suggestions related to this study. First, the research questions are answered using the data from the analysis to provide a better understanding of the results. Then, suggestions for improving the GPE and each of its components are given. The suggestions are followed by a discussion of teaching implications of this study. Finally, suggestions for further research are offered, after which a final conclusion is presented.

Discussion of Research Questions

I. Does the overall content of the German Proficiency Exam represent the general language ability?

It is crucial for a general proficiency test to adequately cover the most important components of language ability. The analysis performed showed very positive results. The content of the German Proficiency Exam with its seven components covers all the language ability construct areas suggested by Bachmann (1996) in his model of general language ability. However, it is important to remember that this definition of language ability represents only a general language construct, and does not provide specific definitions about the construct of separate language skill areas, such as listening, speaking, reading, writing, grammar, and

vocabulary. The following are general suggestions concerning each of the components of the GPE to increase validity:

- Suggestions
    - Define the construct of each GPE component and base the content of the test on the specific areas of the defined language skill constructs.
    - On the basis of the defined language skill construct, develop or adapt existing proficiency level definitions.
    - Provide clearly defined scoring procedures with a detailed scoring key based on the proficiency levels for each component.
    - Include more scorer training concerning the proficiency levels, the scoring procedures and the use of the scoring key.

II. How reliable is each component of the German Proficiency Exam?

The reliability for the listening comprehension exam, the reading exam, the grammar exam, the strong verbs exam, and the vocabulary exam of the GPE was estimated with the Cronbach's alpha reliability coefficient. The reliability values of each of these components are 0.75, 0.89, 0.88, 0.92, and 0.94 respectively. The reliability coefficient of the grammar, reading, and listening component are lower than expected, and coefficient alpha of the strong verbs and vocabulary component are very high. Usually, the reliability level of test scores is influenced by the number of items on a test and the spread of proficiency levels. The more items a test has the more reliable the test scores are. If the range of proficiency levels of students is small and skewed towards the higher scores, the reliability values are higher also. However, even if the test scores are behaving reliably, it does not mean that the test

scores are valid, as we can see when taking a closer look at the components' items. Eight different ways to increase reliability of test scores were presented in Chapter Two, pp. 34-38, which should be considered to make the GPE scores more reliable.

It was not possible to calculate a reliability coefficient for the writing and speaking component, since students received a single score for each component. Further, it was not possible to estimate inter-rater reliability because the components were either scored by one scorer or the scores given by multiple scorers were not recorded. It is crucial for a test, such as the writing and speaking component, to be able to estimate the inter-rater reliability so that conclusions can be drawn about the reliability of test scores. The information gained through inter-rater reliability estimates can help to improve such tests.

Specific suggestions to improve the reliability of the test scores of each component will be given in the following sections.

III. Research questions and discussion for each of the German Proficiency Exam components:

Listening Comprehension

a) How difficult are the items in relation to one another?

b) How well do the items discriminate between the different proficiency levels of students?

The item facility of each item ranging from 0.89 to 0.97 showed that all the items seem to be very easy. The item facility indicates that the students receive a very high score for each item. There can be a couple of reasons for that: either the items

themselves are too easy for the listening ability of the students or the scoring procedures are not clearly defined and only a limited range of scores is given.

The discrimination value for six of the eleven items is below 0.19 and the value of five items lies between 0.26 and 0.27. Therefore, all the items of the listening component do not discriminate well between stronger and weaker students. This has a negative effect on the validity of this component. If a test cannot discriminate well between proficiency levels, the scores of that test can be giving misleading and incorrect information. The reason for low discrimination values in this case is that the items seem not to be very difficult for students. Both stronger and weaker students answer correctly, which does not differentiate between the various levels of listening skill ability.

c) Is there sufficient variation in test scores?

Both the median and mode of all eleven items is 10, indicating that there is minimal variation in test scores. As mentioned above, the items are either very easy or the full scoring range is not being used.

- Suggestions
    - Add more listening passages of medium length and higher difficulty and ask several related questions for each of the listening passages. That will increase the number of items, improving the item facility and discrimination, thereby positively influencing the reliability of test scores.
    - Provide a detailed definition of the scoring criteria and develop a detailed scoring key.

Reading

a) How difficult are the items in relation to one another?

b) How well do the items discriminate between the different proficiency levels?

The item difficulty ranges from 0.57 to 0.87, which indicates a good range of difficulty with a couple of harder items and a few moderate items. Items 3, 4, 6, and 7 discriminate well between proficiency levels. Items 1, 2 and 5 discriminate moderately between proficiency levels and should reviewed for revision. Overall, the items are a good indication of the reading ability of students.

c) Is there sufficient variation in test scores?

Even though the range of test scores for each item is very wide, the scores of items 1 and 2 concentrate around the score 8 and 9, which has an effect on the discrimination and facility of the item. For these items there is not sufficient variation in test scores.

Even though the reading component only has seven items and the variation of test scores for two items is limited, the component overall does show a relatively good reliability coefficient and can be considered a good assessment tool for the language skill of reading.

- Suggestions
    - Review items 1, and 2 to determine whether the difficulty of these items should be increased to increase discrimination ability.
    - Consider item 5 for revision to improve the discrimination value.
    - More specific questions for each of the reading passages could be asked.

Grammar

a) How difficult are the items in relation to one another?

b) How well do the items discriminate between the different proficiency levels of students?

About two thirds of the items of the grammar component can be considered easy. The remaining third of items is divided between moderate and difficult items. This trend of easy items can be caused by the method of testing, which in this case is all 'fill-in-the-blank' and might cover limited grammar functions.

The results of the item facility and item discrimination analyses showed that there are many easy items in general, which seem to lower the discrimination ability of the items. A few items were extremely difficult and need to be examined further, especially items that are both difficult and have low discrimination values, such as items 31, 34, 41, 46, and 54. Low item facilities with low discrimination values indicate that an item may be ambiguous. Most of the easy items do not discriminate well. Most of the 17 moderately difficult items discriminate well.

The department should consider modification or replacement of the easier and less discriminating items. These items could be changed to be harder and cover language material that is more appropriate for higher levels of grammatical knowledge. In order to do that, it is important to be aware of all the components of a grammar construct and to specify which grammar components and functions are more prevalent in higher levels of language proficiency.

- Suggestions
    - Revise items 31, 34, 41, 46, and 54 to be easier

- Review items 1, 5, 6, 8, 9, 10, 12, 13-19, 22, 28-31, 33, 34, 36, 45, 46, 47, 50, 51, and 54 to take them off the test or revise them to make them harder.
- It should be considered that a variety of methods should be used for assessing grammar, so that different language functions can be tested and a variety of grammar functions can be covered. Other methods can include rewriting sentences using a specific grammar function, constructing sentences using sentence parts provided in their basic form, writing short sentences containing a specific grammar function in response to a clue sentence given, and substituting sentence parts with an alternate grammar function.

Strong Verbs

a) How difficult are the items in relation to one another?
b) How well do the items discriminate between the different proficiency levels of students?

The item facility and item discrimination of the strong verbs items show that all the grammatical forms of verb number 7 and verb number 9 discriminate well and have a good item facility value. These values, however, are misleading if we consider the scoring procedure that was explained in the previous two chapters. The examinees might have known the grammatical form, but choose the wrong verb from the clue given. Most of the items of the grammatical forms asked for in columns A, B, E, H, I, and J show very low discrimination (except items of verb 7 and 9), and are thus not very useful. The method used to assess these grammatical forms might

not be challenging enough, and should be considered for revision. Methods of assessing grammatical functions of strong verbs should include a context in which these forms of strong verbs are used. These might make easy grammatical forms more challenging and would avoid guessing the strong verb from an ambiguous clue. Columns C, D, F, and G contain more items that are discriminating better, which points out that the grammatical forms asked for in these columns are more challenging.

- Suggestions
    - The method for assessing the grammatical form indicated in columns A, B, E, H, I, and J could be changed to be more challenging.
    - Consider more complex methods for assessing strong verbs. Grammatical forms of strong verbs could be assessed in context of a sentence or a short paragraph. That might avoid guessing the wrong verb from an ambiguous clue (as is the case in rows 7 and 9)
    - The scoring procedure of this component should be revised. Two points for each item could be given, one for the correct verb and one for correct grammatical form.

Vocabulary

a) How difficult are the items in relation to one another?

Most of the items are extremely hard or very easy; there are only a few items with moderate facility. A component containing many words that are not commonly used or are antiquated can contribute to low item facility. These items should be revised and possibly substituted by more common words. In addition, the scoring procedure

also has a negative effect on the facility values. If no credit is given for knowing the partial meaning of a word, the scores of the vocabulary component do not reflect how much vocabulary an individual has. It is crucial to revise the scoring procedure of this component to make the scores more valid.

b) How well do the items discriminate between the different proficiency levels of students?

The discrimination value of 28 items lies below 0.19 and should either be rejected or revised. The discrimination ability of 11 items is marginal with a value between 0.20 and 0.29, suggesting that they are in need of some improvement. 15 items discriminate reasonably well with a value between 0.30 and 0.39, but could possibly be improved. There are 46 items remaining, which seem to be very good items. However, due to the scoring procedure we cannot know the true discrimination ability of the items, since the facility value influences the discrimination ability. In addition, the behavior of the distractors heavily influences both facility and discrimination.

c) How well do the distractors for each item function?

As mentioned in chapter four, making distractors with high facility values harder and making distractors with low facility values easier can make them more plausible and might help to increase the discrimination value in general. Test developers should make sure that all the distractors are plausible. If one distractor is obviously ridiculous, that distractor is not helping to test and discriminate between students. Distractors that contain a correct meaning, but are chosen very often are

not good distractors and need to be changed. Further, an incorrect distractor that is more prominent than the correct distractors needs to be reviewed.

- Suggestions
    - Review and revise items with very low or very high item facility and very low discrimination ability.
    - Revise distractors that don't attract many responses and are thus not plausible at all.
    - Revise distractors that are incorrect but attract more responses than the correct distractors.
    - Revise the scoring procedure of the component. Either have only multiple-choice items with one correct distractor, or give one point for each correct distractor. The total score of the test would be the sum or correct distractors.
    - Review the content of the component for words and distractor meanings that are not commonly used or antiquated and replace them.

Writing

a) How similar are the task options in terms of task difficulty?

The ANOVA and the post-hoc analysis showed that topic 1 is much different than topic 2 and somewhat different than topic 3. Topics 2 and 3 are more similar in terms of task difficulty. That shows that students that choose topic 1 have a disadvantage, because that topic is more difficult. It is necessary to offer topic choices with similar difficulty to avoid disadvantages over other topics. For that reason the topic options should be revised.

b) Is there sufficient variation in test scores?

There is not sufficient variation in test scores, since the scores of each scoring area concentrates around higher scores. This suggests that the rating of individual writing score areas is not differentiating very well among students. As a result, individual area scores may carry little meaning and may not be a valuable source of feedback to students. Also, these scores do not provide any useful feedback to teachers about student ability in the individual scoring area. There are two reasonable explanations for the low variation in test scores. First, there are no detailed proficiency levels specified. And scoring procedures are not defined and no scoring key is provided that can function as a guideline.

In addition, it would be beneficial to have two scorers score the essays and record the two individual scores. Using the scores from the two scorers, inter-rater reliability can be estimated. That information can be used to improve the rating process.

- Suggestions
    - Revise the topics or have only two options so that there is no advantage of one over the other topic.
    - Revise the wording of the questions to make them more specific and clear.
    - Define scoring procedures and provide a detailed scoring key.
    - Double-rate the essays and record the two scores, so that inter-rater reliability can be established.

Speaking

a) Is there sufficient variation in test scores?

As with the writing component, there is not much variation in test scores for the grammar or the pronunciation parts of the speaking test. The descriptive statistics and the bar graphs for the grammar and pronunciation all show a very high mean and tend to cluster around high scores. If the full range of scores is not used, it is not possible to identify the range of ability of the group of students and the scores of each individual student cannot provide useful feedback about their speaking ability. As mentioned previously, there are several possible reasons for that. First, the group of scorers may not be very familiar with the proficiency levels and scoring key and may use a limited portion of the scoring range. Alternatively, one question that the students discuss during their speaking exam might not provide enough information about their speaking ability. Several shorter questions about a couple different topics might provide more information, not only about their grammar and pronunciation ability, but also about other abilities that make up the overall speaking ability. Finally, the German Section focuses on teaching its students good speaking abilities, and the students have reached a high level of speaking proficiency by the end of their course of study.

A way to still be able to identify the lower levels of proficiency, within a group of high proficiency is to describe and define detailed levels within the higher proficiency levels. That way, teachers can discriminate the weaker students from the stronger students and also provide more detailed information and feedback to their students about their speaking ability.

- Suggestions
    - Divide the higher proficiency levels into several separate proficiency levels.
    - Develop a more detailed scoring key on the basis of the specified proficiency levels.
    - Train all the scorers so that they are familiar with the scoring procedures, including proficiency levels and scoring key.
    - Double-rate the exam, if possible, and record the individual scores to estimate inter-rater reliability.

Pedagogical Implications

A well-defined language ability construct, either adapted from an existing theory of language ability or tailored to the needs of the department, used as the basis for developing a language test can have a very positive effect on teaching. Common proficiency levels for each language skill that are defined on the basis of the language construct can guide the development of objectives for all the language courses of the German program. Language instructors can have more clarity on how the courses connect from the first beginning language class to the last advanced course and can have more guidance for teaching the individual courses. Well developed tests with a high score validity can provide students with meaningful feedback on their proficiency in the language skills. They can use that feedback to concentrate on certain aspects of language skills and improve their overall language proficiency. In addition, valid and reliable test scores can provide teachers with useful information

regarding language skills and areas of teaching that might need improvement. This can have a positive washback effect on the teaching and learning of language.

Suggestions for Future Research

Compared to the vast amount of available evidence, this study collected only a limited amount of validity related evidence using a few methods of data analysis. The process of validity is not a one-time treatment, but an ongoing process, making it necessary to continually collect evidence that supports the interpretation and use of the test. Therefore, there is an almost unlimited amount of evidence available for getting more and different information about the validity of the test.

First, the suggestions for improving the GPE should be applied. Then, a reliability coefficient should be estimated in order to investigate whether the changes have a positive effect on the reliability of test scores.

Second, since this study could only cover a limited amount of evidence in the time provided, it investigated the general language ability and provided one model of language construct. More information about the validity of the test scores could be gained from investigating each specific language skill construct and analyzing the content of each GPE component for coverage of the specific language skill construct areas and functions.

Third, the total scores of the GPE could be correlated and compared with each individual total score of the component to investigate how similar each component is in terms of difficulty and how much the component scores correlate with the total score.

In addition, a deeper analysis of the scoring procedures would be useful to improve the process of scoring. Inter-rater reliability could be established for the components scored by multiple raters, such as the writing and speaking components.

Besides collecting quantitative evidence, more qualitative evidence could be collected and analyzed. The examinees could be interviewed or questionnaires could be filled out about the experience of taking, administering, or scoring the GPE. More insight could be gained about how the examinees would use the results of the GPE in their future career and in what ways it does or does not help them to improve their German language skills.

Finally, it could be investigated where to set a cut score. Then, it should be analyzed if that cut score is set at an effective and useful level. A cut score has the potential to be very beneficial to the proficiency exam. A cut score would allow and motivate students to improve performance and would ensure that the students leaving the German program are proficient enough to use the language in their careers.

Conclusions

The purpose of this study was to examine the validity of the GPE scores using quantitative and qualitative research methods to help improve the exam. The findings of this study are not intended to be generalized to a larger population. Rather, this research functions as a way to give more information about a specific testing situation. Nevertheless, the theories discussed in this research can be applied for the improvement of any language test situation.

This study provided specific insight into the GPE and gave information about some aspects of validity of the GPE. Overall, the GPE provides a good basis as a proficiency test of German for the German Section at BYU. Numerous suggestions were given to make the test scores more reliable and valid. As discussed, it would be beneficial to clearly define language ability and specify language skill constructs. Clearly defined proficiency levels for each skill area and detailed scoring keys could function as a guideline for an effective scoring process. In addition, clear definitions of proficiency levels would also provide a means to give more meaningful feedback to the students. Scorers who are trained to be more familiar with the proficiency levels and the scoring procedures can contribute to more reliable scores.

The value of the German Proficiency Exam in showing valid and reliable test scores is very important for the German Section at Brigham Young University and their students. The purpose of the GPE is to provide information to students, teachers, administrators of the department, and future employers. It is crucial that the results of the GPE represent the true proficiency of the students so that the decisions made based on the GPE scores can have a positive influence on society. Since validity is a continuous process, validation studies should be continuously performed in order to build a larger and larger base of evidence.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, UK: Oxford University Press.

Bachman, L. F., (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L. F., Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Brigham Young University (2006). *Accreditation self-study report.* Retrieved April 12, 2008 from http://accredit.byu.edu/.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* New York, NY: The McGraw-Hill Companies.

Brown, J. D. (1989). Language program evaluation: A synthesis of existing possibilities. In K. Johnson (Ed.) *The Second Language Curriculum* (p. 222 – 241). London, UK: Cambridge University Press.

Brown, J. D., Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge, UK: Cambridge University Press.

Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.

Council of Europe (2004). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Davies, A. (1999). *Disctionary of language testing*. Cambridge, UK: Press Syndicate of the University of Cambridge.

Fulcher, G. (2003). Defining the construct. In C. N. Candlin (Ed.), *Testing second language speaking* (pp. 18-49). Harlow, UK: Pearson Longman.

Germanic & Slavic Languages (2007). Expected learning outcomes, evidence and assessment. Retrieved March 24, 2008 from https://learningoutcomes.byu.edu/wiki/index.php/Germanic_and_Slavic_Languages.

Hudson, T. (1989). Mastery decisions in program evaluation. In K. Johnson (Ed.) *The Second Language Curriculum* (p. 259 – 269). London, UK: Cambridge University Press.

Hughes, A. (2003). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.

Hughes, A., Porter, D. and Weir, C. J. (1988). *Validating the ELTS test: A critical review*. Cambridge, UK: The British Council and the University of Cambridge Local Examination Syndicate.

Kunnan, A. J. (1998). Approaches to validation in language assessment. In Kunnan, A. J. (Ed.), *Validation in language assessment: Selected papers from the 17th language testing colloquium, Long Beach* (pp. 1-16), Mahwah, NJ: Lawrence Erlbaum Associates.

Lado, R. (1961). *Language Testing*. New York: McGraw-Hill.

Linn, R.L. and Gronlund, N.E. (2000). Measurement and Assessment in Teaching (8th ed.). Upper Saddle River, NJ: Merrill.

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York, NY: Macmillan.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18, 5-11.

Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of Educational Research* (6th ed.), New York, NY: Macmillan.

Palmer, A. (1991). The role of language testing in language program evaluation. In Anivan, S. (Ed.), *Issues in Language Programme Evaluation in the 1990's*. Anthology Series 27.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

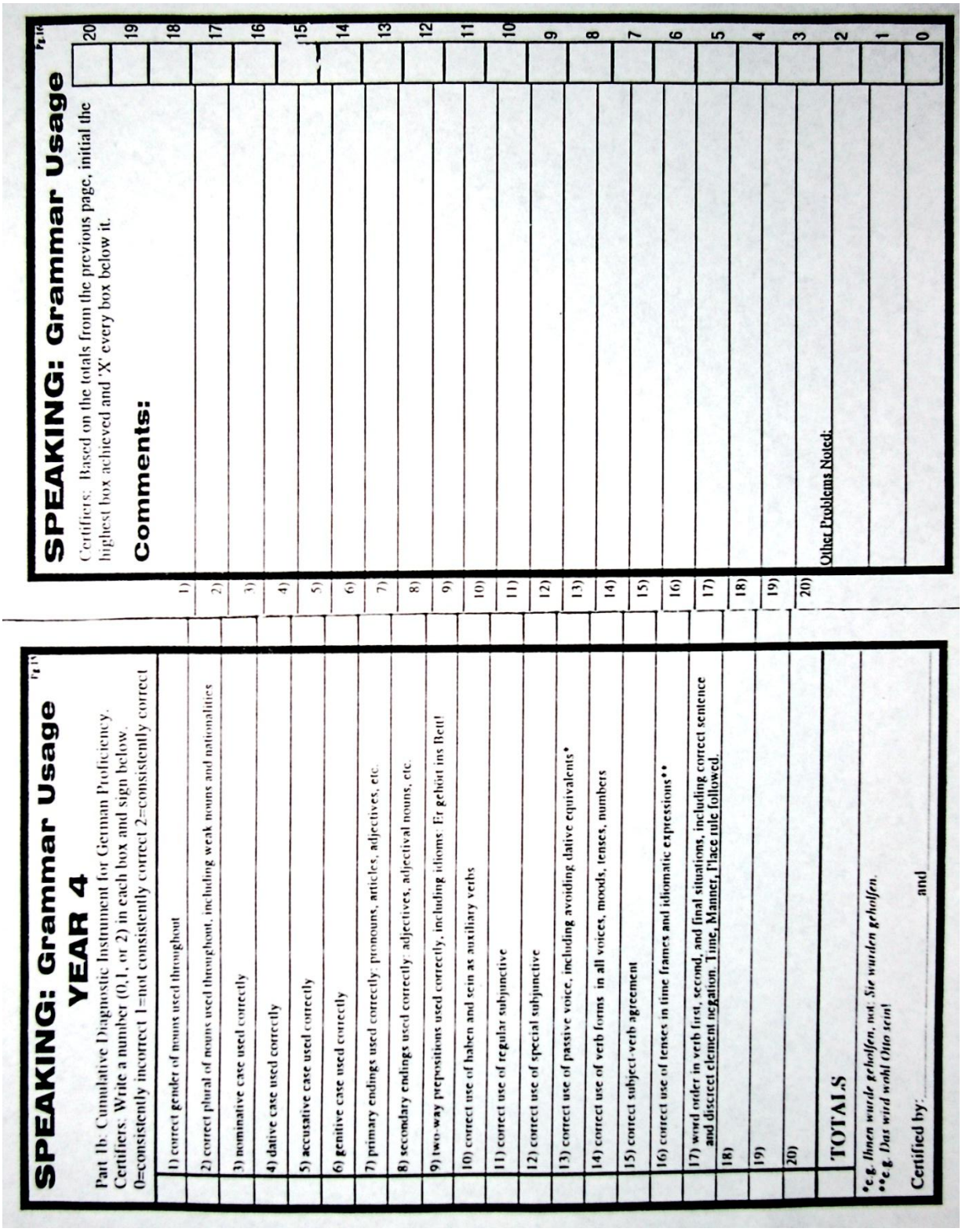
Tanner, J. S. (2006). *Building a better house of learning.* Address delivered at BYU Annual University Conference faculty session Aug 29, 2006. Retrieved June 17, 2007 from http://ctl.byu.edu/?page_id=11.

Weir, C. J. (2005). *Language testing and validation: an evidence-based approach*. New York, NY: Palgrave Macmillan.

Woodford, P. (1980). Foreign language testing. *The Modern Language Journal,* 64(1), 97-102.

APPENDIX A

Speaking: Grammar Usage Diagnostic Instrument



SPEAKING: Grammar Usage YEAR 4

Part 1b. Cumulative Diagnostic Instrument for German Proficiency.
Certifiers: Write a number (0,1, or 2) in each box and sign below.
0=consistently incorrect 1=not consistently correct 2=consistently correct

	1) correct gender of nouns used throughout	1)
	2) correct plural of nouns used throughout, including weak nouns and nationalities	2)
	3) nominative case used correctly	3)
	4) dative case used correctly	4)
	5) accusative case used correctly	5)
	6) genitive case used correctly	6)
	7) primary endings used correctly: pronouns, articles, adjectives, etc.	7)
	8) secondary endings used correctly: adjectives, adjectival nouns, etc.	8)
	9) two-way prepositions used correctly, including idioms: Er geht ins Bett	9)
	10) correct use of haben and sein as auxiliary verbs	10)
	11) correct use of regular subjunctive	11)
	12) correct use of special subjunctive	12)
	13) correct use of passive voice, including avoiding dative equivalents*	13)
	14) correct use of verb forms in all voices, moods, tenses, numbers	14)
	15) correct subject-verb agreement	15)
	16) correct use of tenses in time frames and idiomatic expressions**	16)
	17) word order in verb first, second, and final situations, including correct sentence and discreet element negation. Time, Manner, Place rule followed.	17)
	18)	18)
	19)	19)
	20)	20)
	TOTALS	

*e.g. Ihnen wurde geholfen, not: Sie wurden geholfen.
**e.g. Das wird wohl Otto sein!

Certified by: _____ and _____

SPEAKING: Grammar Usage

Certifiers: Based on the totals from the previous page, initial the highest box achieved and 'X' every box below it.

Comments:

	20
	19
	18
	17
	16
	15
	14
	13
	12
	11
	10
	9
	8
	7
	6
	5
	4
	3
	2
	1
	0

Other Problems Noted:

APPENDIX B

Speaking: Pronunciation Diagnostic Instrument

SPEAKING: Pronunciation

YEAR 4

Part Ia Cumulative Diagnostic Instrument for Authentic German Pronunciation Certifiers: Under the appropriate Year of Study write a number (0,1, or 2) in each box and sign below
0=consistently incorrect 1=not consistently correct 2=consistently correct

1) no American r	
2) no American l	
3) clear distinction between [^] and uvular [R] (Tier, Tiere)	
4) good, rounded ü	
5) good, rounded ö	
6) clear short-long and tense-lax distinctions (beten-Betten)	
7) good [X], no confusion with [ç] or [k]	
8)good [ç], no confusion with [X] or [k]	
9) accent on correct syllable	
10) no false diphthongs (air leaped goat)	
11) strong [ts] (zehn)	
12) correct ʃ (Schuh, Spiel, schwer)	
13) correct [z] (senden, Hasen)	
14) correct [s] (Slave)	
15) correct [pf] (Pfund)	
16) correct [kn] (Knie)	
17) correct [gn] (Gnade)	
18) correct [kv] (quer)	
19) correct unvoicing at syllable beak: [ta:k], [li:plos]	
20) assimilation of final syllables haben=[ha:bm], mögen=[mø:gn], etc.	
TOTALS	

Certified by: ... and ...

SPEAKING: Pronunciation

Certifiers: Based on the totals from the previous page, initial the highest box achieved and 'X' every box below it.

Comments:

1)	
2)	
3)	
4)	
5)	
6)	
7)	
8)	
9)	
10)	
11)	
12)	
13)	
14)	
15)	
16)	
17)	
18)	
19)	
20)	

20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0

Other Problems Noted:

APPENDIX C

Expanded Description of Proficiency Level

Range: Approx. 16-20 Able to use the language with sufficient accuracy to participate effectively in most formal and informal discourse on practical, social, professional, and abstract topics. Can discuss special fields of competence and interest with ease. Can support opinions and hypothesize, but may not be able to tailor language exactly to audience or discuss in depth highly abstract or unfamiliar topics. May only be partially familiar with regional or other dialect variants. Commands a wide variety of interactive strategies and shows awareness of discourse strategies involving the ability to distinguish main ideas from supporting information through syntactic, lexical, and suprasegmental features such as pitch, stress, intonation. Sporadic errors may occur, particularly in low frequency structures and some complex high-frequency structures such as those common to formal writing, but no patterns of error are evident. Errors do not disturb the native speaker/reader or interfere with communication.

Range: Approx. 11-15 Able to satisfy the requirements of everyday situations and routine school and work requirements. Can handle with confidence but not with facility complicated tasks and social situations, such as elaborating, complaining, and apologizing. Can narrate and describe with some details, linking sentences together smoothly. Can communicate facts and talk casually about topics of current public and personal interest, using general vocabulary. Shortcomings can often be smoothe over by communicative strategies such as pause filler, stalling devices, and different rates of speech. Circumlocution which arises from vocabulary or syntactic limitations very often is quite successful, though some groping for words may still be evident. Can be understood without difficulty by native interlocutors, though with some misunderstandings arising. Errors are evidence and cause struggles for appropriate forms.

Range: Approx. 6-10 Able to handle successfully a variety of uncomplicated, basic, and communicative tasks and social situations. Can talk simply about self and family members. Can ask and answer questions and participate in simple conversations on topics such as personal history and leisure time activities. Speech may be characterized by frequent long pauses, since the smooth incorporation of even basic discourse strategies is often hindered as the student struggles to create – also in writing – appropriate language forms. Pronunciation may continue to be strongly influenced by first language and fluency may still be strained. Although misunderstandings arise, can be generally understood by sympathetic interlocutors, especially with much repetition. Strong interference from native language is evident. Can ask and answer questions and respond to simple statements, although in a highly restricted manner and with much linguistic inaccuracy, though ok for basic needs.

Range: Approx. 0-5 Able to satisfy partially the requirements of basic communicative exchange by relying heavily on learned utterances but occasionally expanding these through simple recombinations of their elements. Can ask questions or make statements involving learned material. Shows signs of spontaneity although this falls short of real autonomy of expression. Speech continues to consist of learned utterances rather than of personalized, situationally adapted ones. Vocabulary centers on areas such as basic objects, places, and most common kinship terms. Pronunciation is still strongly influenced by first language. Errors are frequent and, in spite of repetition, cause difficulty being understood even by the most sympathetic interlocutors. Often can be understood only with great difficulty. Sometimes even the simplest utterances fail. Production consists of isolated words and perhaps a few high-frequency phrases. Essentially no functional communicative ability.

APPENDIX D

Guidelines for Evaluating Proficiency Exam Orals

*Please use the sheets given you to track the student's strengths and weaknesses in oral proficiency, and feel free to write any comments on the sheets. We recommend that one member of the team track the Grammar Usage, while the other tracks Pronunciation.

*Please evaluate each student using a 20 point scale. We are including a description of the proficiency levels for your reference. The scores should be distributed according to this approximate standard:

20	
19	Sehr gut [A]
18	
17	
16	But [B]
15	
14	Befriedigend [C]
13	
12	Ausreichend [D]

APPENDIX E

Summary Score Sheet

Listening	Reading	Pronunc.	SpokenGr	Writing	Grammar	Strong Verbs	Vocabulary	%	ZDaf	ZMP	ZOP	OPI	Class
							Thousands of words in vocab						
							20k	100	sg	sg	sg	S	
							19k	95	sg	sg	sg	S	
							18k	90	sg	sg	g	A+	
							17k	85	sg	sg	g	A+	
							16k	80	sg	sg	be	A	
							15k	75	sg	g	be	A	
							14k	70	sg	g	aus	IH	
							13k	65	sg	be	aus	IH	
							12k	60	g	be	nb	IH	
							11k	55	g	aus	nb	IM	
							10k	50	be	aus	nb	IM	
							9k	45	be	nb	nb	IM	
							8k	40	aus	nb	nb	IL	
							7k	35	aus	nb	nb	IL	
							6k	30	nb	nb	nb	NH	
							5k	25	nb	nb	nb	NH	
							4k	20	nb	nb	nb	NM	
							3k	15	nb	nb	nb	NM	
							2k	10	nb	nb	nb	NL	
							1k	5	nb	nb	nb	NL	

Key abbreviations:
sg=sehr gut (very good)
g=gut (good)
be=bestanden (passed)
aus=ausreichend (adequate)
nb=nicht bestanden (failed)
NL=Novice-Low
NM=Novie-Mid
NH=Novice-High
IL=Intermediate-Low
IM=Intermediate-Mid
IH=Intermediate-High
A=Advanced
A+=Advanced-Plus
S=Superior

APPENDIX F

Vocabulary Item and Distractor Analysis

Item	IF	Item Discr.	Distr.	Distractor *p*	Item	IF	Item Discr.	Distr.	Distractor *p*
1	0.31	0.15	A	0.31	10	0.68	0.24	A	0.68
			B	0.21				B	0.06
			C	0.16				C	0.18
			D	0.31				D	0.09
2	0.02	0.05	A	0.22	11	0.35	0.73	A	0.07
			B	0.04				B	0.01
			C	0.97				C	0.36
			D	0.35				D	0.60
3	0.02	0.05	A	0.13	12	0.73	0.51	A	0.02
			B	0.91				B	0.77
			C	0.00				C	0.20
			D	0.09				D	0.04
4	0.05	0.12	A	0.79	13	0.61	0.49	A	0.62
			B	0.09				B	0.16
			C	0.03				C	0.12
			D	0.20				D	0.12
5	0.00	0.00	A	0.06	14	0.00	0.00	A	0.08
			B	0.72				B	0.07
			C	0.35				C	0.28
			D	0.17				D	0.78
6	0.49	0.22	A	0.03	15	0.85	0.24	A	0.06
			B	0.05				B	0.92
			C	0.67				C	0.02
			D	0.45				D	0.08
7	0.27	0.32	A	0.33	16	0.20	0.42	A	0.55
			B	0.52				B	0.30
			C	0.09				C	0.58
			D	0.32				D	0.42
8	0.33	0.14	A	0.07	17	0.24	0.47	A	0.62
			B	0.33				B	0.44
			C	0.26				C	0.10
			D	0.35				D	0.22
9	0.78	0.25	A	0.07	18	0.65	0.36	A	0.24
			B	0.09				B	0.09
			C	0.84				C	0.74
			D	0.06				D	0.04

Item	IF	Item Discr.	Distr.	Distractor *p*	Item	IF	Item Discr.	Distr.	Distractor *p*
19	0.77	0.22	A	0.09	29	0.89	0.15	A	0.91
			B	0.04				B	0.08
			C	0.12				C	0.02
			D	0.85				D	0.00
20	0.66	0.54	A	0.05	30	0.65	0.47	A	0.20
			B	0.01				B	0.04
			C	0.28				C	0.13
			D	0.68				D	0.68
21	0.82	0.31	A	0.06	31	0.56	0.46	A	0.59
			B	0.84				B	0.42
			C	0.08				C	0.00
			D	0.06				D	0.02
22	0.92	0.15	A	0.97	32	0.61	0.64	A	0.09
			B	0.01				B	0.71
			C	0.06				C	0.09
			D	0.01				D	0.20
23	0.38	0.69	A	0.59	33	0.53	0.56	A	0.39
			B	0.41				B	0.06
			C	0.02				C	0.56
			D	0.01				D	0.02
24	0.94	0.15	A	0.02	34	0.95	0.07	A	0.03
			B	0.02				B	0.02
			C	0.03				C	0.01
			D	0.97				D	0.96
25	0.79	0.29	A	0.82	35	0.42	0.66	A	0.44
			B	0.09				B	0.31
			C	0.10				C	0.15
			D	0.00				D	0.14
26	0.56	0.41	A	0.27	36	0.88	0.22	A	0.05
			B	0.57				B	0.01
			C	0.08				C	0.93
			D	0.09				D	0.06
27	0.63	0.42	A	0.09	37	0.27	0.56	A	0.05
			B	0.04				B	0.46
			C	0.72				C	0.29
			D	0.23				D	0.28
28	0.59	0.44	A	0.21	38	0.64	0.68	A	0.07
			B	0.63				B	0.08
			C	0.08				C	0.20
			D	0.12				D	0.65

Item	IF	Item Discr.	Distr.	Distractor *p*	Item	IF	Item Discr.	Distr.	Distractor *p*
39	0.80	0.42	A	0.09	49	0.32	0.47	A	0.23
			B	0.07				B	0.05
			C	0.03				C	0.42
			D	0.82				D	0.34
40	0.45	0.39	A	0.48	50	0.73	0.47	A	0.07
			B	0.32				B	0.18
			C	0.06				C	0.87
			D	0.15				D	0.06
41	0.61	0.64	A	0.21	51	0.33	0.68	A	0.36
			B	0.11				B	0.36
			C	0.63				C	0.22
			D	0.07				D	0.08
42	0.48	0.53	A	0.53	52	0.80	0.41	A	0.15
			B	0.05				B	0.03
			C	0.42				C	0.01
			D	0.05				D	0.83
43	0.44	0.47	A	0.45	53	0.32	0.69	A	0.31
			B	0.27				B	0.06
			C	0.09				C	0.30
			D	0.21				D	0.35
44	0.38	0.75	A	0.26	54	0.66	0.31	A	0.70
			B	0.13				B	0.04
			C	0.24				C	0.14
			D	0.42				D	0.14
45	0.68	0.63	A	0.09	55	0.62	0.32	A	0.12
			B	0.72				B	0.02
			C	0.17				C	0.62
			D	0.06				D	0.22
46	0.49	0.66	A	0.50	56	0.32	0.53	A	0.48
			B	0.25				B	0.04
			C	0.23				C	0.42
			D	0.08				D	0.33
47	0.57	0.53	A	0.24	57	0.41	0.61	A	0.53
			B	0.05				B	0.07
			C	0.57				C	0.01
			D	0.14				D	0.49
48	0.61	0.41	A	0.73	58	0.31	0.25	A	0.18
			B	0.21				B	0.54
			C	0.06				C	0.03
			D	0.16				D	0.54

Item	IF	Item Discr.	Distr.	Distractor *p*	Item	IF	Item Discr.	Distr.	Distractor *p*
59	0.54	0.66	A	0.63	69	0.60	0.37	A	0.36
			B	0.28				B	0.03
			C	0.04				C	0.78
			D	0.15				D	0.02
60	0.17	0.47	A	0.26	70	0.70	0.56	A	0.80
			B	0.27				B	0.26
			C	0.18				C	0.04
			D	0.54				D	0.02
61	0.42	0.47	A	0.31	71	0.15	0.44	A	0.89
			B	0.13				B	0.18
			C	0.43				C	0.21
			D	0.15				D	0.42
62	0.80	0.44	A	0.82	72	0.65	0.27	A	0.65
			B	0.09				B	0.09
			C	0.08				C	0.13
			D	0.02				D	0.12
63	0.92	0.19	A	0.03	73	0.38	0.42	A	0.58
			B	0.92				B	0.69
			C	0.03				C	0.06
			D	0.00				D	0.12
64	0.32	0.34	A	0.32	74	0.42	0.53	A	0.31
			B	0.20				B	0.23
			C	0.39				C	0.42
			D	0.07				D	0.06
65	0.75	0.19	A	0.17	75	0.26	0.53	A	0.28
			B	0.02				B	0.21
			C	0.83				C	0.19
			D	0.03				D	0.32
66	0.16	0.49	A	0.51	76	0.12	0.31	A	0.13
			B	0.63				B	0.36
			C	0.28				C	0.18
			D	0.07				D	0.45
67	0.57	0.58	A	0.63	77	0.54	0.39	A	0.02
			B	0.26				B	0.56
			C	0.09				C	0.20
			D	0.10				D	0.26
68	0.13	0.31	A	0.21	78	0.09	0.25	A	0.20
			B	0.37				B	0.15
			C	0.32				C	0.68
			D	0.38				D	0.12

Item	IF	Item Discr.	Distr.	Distractor *p*	Item	IF	Item Discr.	Distr.	Distractor *p*
79	0.22	0.41	A	0.26	89	0.17	0.36	A	0.31
			B	0.01				B	0.21
			C	0.67				C	0.53
			D	0.31				D	0.06
80	0.53	0.53	A	0.32	90	0.39	0.00	A	0.43
			B	0.06				B	0.08
			C	0.53				C	0.08
			D	0.08				D	0.41
81	0.82	0.15	A	0.86	91	0.72	0.17	A	0.75
			B	0.01				B	0.03
			C	0.13				C	0.02
			D	0.03				D	0.22
82	0.62	0.25	A	0.09	92	0.73	0.19	A	0.11
			B	0.09				B	0.10
			C	0.18				C	0.04
			D	0.64				D	0.74
83	0.27	0.36	A	0.28	93	0.25	0.15	A	0.07
			B	0.55				B	0.30
			C	0.03				C	0.32
			D	0.16				D	0.37
84	0.58	0.31	A	0.06	94	0.01	0.02	A	0.12
			B	0.60				B	0.20
			C	0.17				C	0.43
			D	0.21				D	0.28
85	0.21	-0.05	A	0.07	95	0.54	0.00	A	0.09
			B	0.04				B	0.07
			C	0.34				C	0.58
			D	0.65				D	0.29
86	0.51	0.39	A	0.01	96	0.01	0.03	A	0.54
			B	0.36				B	0.22
			C	0.13				C	0.45
			D	0.54				D	0.01
87	0.01	0.03	A	0.18	97	0.36	-0.02	A	0.45
			B	0.75				B	0.24
			C	0.10				C	0.32
			D	0.15				D	0.14
88	0.63	0.12	A	0.21	98	0.78	0.10	A	0.12
			B	0.01				B	0.13
			C	0.14				C	0.01
			D	0.67				D	0.79

Item	IF	Item Discr.	Distr.	Distractor *p*	Item	IF	Item Discr.	Distr.	Distractor *p*
99	0.52	0.17	A	0.12	100	0.48	0.12	A	0.21
			B	0.14				B	0.16
			C	0.54				C	0.13
			D	0.21				D	0.50