Symposium

Language Testing in Austria: Towards a Research Agenda

October 19-20, 2018

Book of Abstracts

https://www.aau.at/en/language-testing-centre/conferences/LTiA2018/

ALPEN-ADRIA UNIVERSITÄT
KLAGENFURT | WIEN GRAZ

Organising committee:
Günther Sigott
Hermann Cesnik
Nikola Dobrić

With special thanks to:

ösd

Language Testing and Evaluation 40

Günther Sigott (Ed./Hrsg.)

Language Testing in Austria: Taking Stock
Sprachtesten in Österreich: Eine Bestandsaufnahme

PETER LANG

Language Testing and Evaluation

Series editors: Rüdiger Grotjahn, Claudia Harsch and Günther Sigott

Volume 40

Zur Qualitätssicherung und Peer Review der vorliegenden Publikation

Die Qualität der in dieser Reihe erscheinenden Arbeiten wird vor der Publikation durch die Herausgeber der Reihe geprüft.

Notes on the quality assurance and peer review of this publication

Prior to publication, the quality of the work published in this series is reviewed by the editors of the series.

Günther Sigott (Ed./Hrsg.)

Language Testing in Austria: Taking Stock
Sprachtesten in Österreich: Eine Bestandsaufnahme

PETER LANG

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

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

ISSN 1612-815X
ISBN 978-3-631-74938-8 (Print)
E-ISBN 978-3-631-75973-8 (E-Book)
E-ISBN 978-3-631-75974-5 (E-PUB)
E-ISBN 978-3-631-75975-2 (MOBI)
DOI 10.3726/b14338

© Peter Lang GmbH
Internationaler Verlag der Wissenschaften
Berlin 2018
All rights reserved..

Peter Lang – Berlin · Bern · Bruxelles · New York ·
Oxford · Warszawa · Wien

All parts of this publication are protected by copyright. Any utilisation outside the strict limits of the copyright law, without the permission of the publisher, is forbidden and liable to prosecution. This applies in particular to reproductions, translations, microfilming, and storage and processing in electronic retrieval systems.

This publication has been peer reviewed.

www.peterlang.com

Contents

Acknowledgements 9

Günther Sigott
Introduction 11

Part I: Test Development

I.1 National Educational Standards Tests

Sandra Eibl, Silke Schwaiger & Simone Breit
Bildungsstandardüberprüfung Deutsch: Kompetenzmodelle und Aufgabenentwicklung 23

Claudia Luger-Bazinger, Ursula Itzlinger-Bruneforth & Claudia Schreiner
Die Umsetzung der kriterialen Rückmeldung in der Bildungsstandardüberprüfung am Beispiel Lesen in Deutsch für die vierte und achte Schulstufe 43

Claudia Mewald, Dave Allan & Andrea Kulmhofer
The development of the Austrian Educational Standards Test for English Listening at grade 8 65

Klaus Siller & Andrea Kulmhofer
The development of the Austrian Educational Standards Test for English Reading at grade 8 85

Claudia Mewald
The development of the Austrian Educational Standards Test for English Speaking at grade 8 109

Andrea Kulmhofer & Klaus Siller
The development of the Austrian Educational Standards Test for English Writing at grade 8 129

I.2 National School-Leaving Exam (Matura)

Jürgen Struger
Deutsch als Unterrichtssprache: Das Konzept der schriftlichen Reife- und Diplomprüfung 155

Hermann Cesnik & Günther Sigott
Die Standardisierte Reife- und Diplomprüfung (SRDP) Deutsch als Unterrichtssprache: Methodenkonzept und Hauptergebnisse der Feldtestung 183

Carol Spöttl, Kathrin Eberharter, Franz Holzknecht, Benjamin Kremmel & Matthias Zehentner
Delivering reform in a high stakes context: from content-based assessment to communicative and competence-based assessment 219

Theresa Weiler & Doris Frötscher
Ensuring sustainability and managing quality in producing the standardized matriculation examination in the foreign languages 241

Fritz Lošek, Peter Glatz, Hermann Niedermayr & Irmtraud Weyrich-Zak
Entwicklung von Testsystemen in den klassischen Sprachen 261

Michael Sörös, Martin Seitz, Renate Glas, Walter Kuchling, Renate Oswald & Andrea Lošek
Aspekte der Implementierung der neuen Testsysteme in den klassischen Sprachen 287

I.3 Tertiary-Level Language Testing

Armin Berger & Helen Heaney
Developing rating instruments for the assessment of Academic Writing and Speaking at Austrian university English departments 325

I.4 German as a Foreign Language

Manuela Glaboniat & Carmen Peresich
Österreichisches Sprachdiplom Deutsch..... 349

Part II: Language Testing Research

II.1 National Educational Standards Tests

Roman Freunberger, Simone Breit & Marcel Illetschko
Beurteilerübereinstimmung und schwer zu beurteilende Texte im Vergleich Training und Bildungsstandardüberprüfung in Deutsch auf der vierten Schulstufe 373

Günther Sigott & Hermann Cesnik
On the discrepancy between rater-perceived and Empirical Item Difficulty. Standardsetting for the receptive skills in the Austrian Educational Standards Tests for English 389

Florian Pibal, Günther Sigott & Hermann Cesnik
The role of error in assessing English Writing in the Austrian Educational Standards Baseline Test 419

Claudia Mewald
The washback of standardised testing in the subject English at lower secondary level 445

Klaus Siller & Ulrike Kipman
Cognitive processes as predictors of item difficulty in the Austrian Educational Standards Baseline Test for English Reading at grade 8 465

II.2 National School-Leaving Exam (Matura)

Doris Frötscher, Michael Themessl-Huber & Theresa Weiler
Analyzing the 2016 standardized matriculation exam results in the foreign languages: post-test analysis 489

Doris Frötscher
Matura washback on the classroom testing of Reading in English 515

Michael Maurer
How do test takers write a blog? A corpus-based study of linguistic features in performances based on the task type Blog 541

Franz Holzknecht
Reading strategies in the standardized Austrian matriculation examination for English 563

Kathrin Eberharter
Validating multiple choice Language in Context items 585

Kathrin Eberharter, Franz Holzknecht & Benjamin Kremmel
Evaluating the effectiveness of a training program for double-raters 607

Matthias Zehentner
Examining the effect of partner choice in peer-to-peer assessment 629

Hermann Cesnik
Annäherung der Transformationsfunktion von Logit Measures auf Fair Measures im Multifacetten-Raschmodell (MFRM) durch logistische Regression. Eine Simulationsstudie 651

II.3 Tertiary-Level Language Testing

Armin Berger
Rating scale validation for the assessment of spoken English at tertiary level 679

Helen Heaney
Operationalizing expeditious reading strategies in tests of English as a foreign language 703

Hans Platzer
A criterion-referenced approach to the Vocabulary Levels Test: validating form B of the 2000 and 3000 word levels, and the Academic Wordlist 723

Acknowledgements

This book is the result of a collaborative effort. I would like to thank all contributors for making this volume possible. Their support and cooperation during two years of intensive work were exceptional, delightful and refreshing. I am grateful to Christina Thamer for her valuable help in the editing process. I would also like to thank Rüdiger Grotjahn and Claudia Harsch for undertaking the time-consuming task of reviewing all thirty-one contributions. My particular thanks go to my long-standing colleague Hermann Cesnik for sharing his vast expertise in statistical methods, psychometrics and beyond, his excellent support in IT matters, his absolute reliability, devotion to the cause, and his friendship throughout the entire project. Thank you, Hermann.

Günther Sigott*

Introduction

This volume marks a decade of modern language testing in Austria. Part I describes developments at all levels of the educational system ranging from elementary school to university level. Part II aims to chart the language testing research territory by bringing together a considerable amount of research that has accompanied and followed national exam development and reform since 2007, thus testifying to the existence of an important and active language testing community in Austria.

1. The situation before the reform

Since 2007 the language testing scene in Austria has changed enormously. In the time before, basic tenets of language test development were, at best, present in the isolated expertise of individual researchers or dedicated language teachers scattered in institutions all over the country. Language testing as an academic discipline was not established and people attempting to establish language testing as a worthwhile field of enquiry within the academic landscape in Austria were seen as exotic. This is not to say that language competences were not assessed. Like in most other countries, there had been a time-honoured tradition of assessing pupils' and students' language achievements towards the end of compulsory schooling at grades 8 or 9, when pupils are between 14 and 15 years of age, in the form of a school certificate. At the end of upper secondary level, when students were at age 18 or 19, their achievements were assessed in a high-stakes examination, the so-called Matura in academic and vocational secondary schools. For those who had chosen to study a philological subject other than their mother tongue, language achievement or proficiency was assessed at the end of their university studies. Questions relating to the reliability or validity of these examinations were rarely asked.

In the first decade of the 21st century, it gradually became clear that the traditional examination procedures were becoming more and more anachronistic. The Common European Framework of Reference (CEFR) for Languages had been around for years and language proficiency requirements and achievements were increasingly being described with reference to the six levels of the CEFR. By 2007,

* Address for correspondence: Günther Sigott, Alpen-Adria-Universität Klagenfurt, guenther.sigott@aau.at.

the situation concerning high-stakes language examinations in Austria gave rise to questions like the following (Sigott 2007:160):

1. Does the exam aim to test the things which it should test?
2. Is it sufficiently clear what the exam is actually testing?
3. Does the exam test the same things in all the schools or university departments?
4. Does / will the exam test the same things every year?
5. Are parallel forms of the exam available?
6. Are results comparable from year to year?
7. Is it clear how the exam relates to the CEFR?
8. If the answer to any of the above questions is YES: What evidence is there?

Answering the first question was difficult because at both the end of secondary level and at the end of tertiary level the desired outcomes had not been described in sufficient detail, presumably because there was a lack of agreement as to what they should be. The second question goes to the heart of test validity, and requires evidence, which, in turn, presupposes a detailed description of the test constructs that constitute the basis for the tests in question. But such evidence was not available in 2007. Given the negative answers to the first two questions, it could safely be assumed that the answer to question three was also negative. After all, if there are no commonly agreed-upon test specifications, coordinating exam content and method is practically impossible. The obvious conclusion was that neither Matura nor B.A. exit examinations were likely to test the same skills or competences across schools or modern language university departments, respectively. For the same reasons, it was unlikely that the test constructs remained stable from year to year, which meant a negative answer to question four as well. In the absence of test specifications and calibrated item banks it was not to be expected that parallel forms of either Matura or B.A. exit examinations existed, which means a negative answer to question five. This indicated that neither exam content nor exam difficulty were likely to be stable over the years in both examination contexts, which suggests a negative answer to question six. Obviously, the prerequisites for attempting to link the examinations to the CEFR were not met and also question seven had to be answered negatively. Given this state of affairs, the conclusion I reached in 2007 read as follows:

> Since the answers to the above questions have been negative, there is no need to approach question eight. In fact, there is no relevant evidence that would justify a positive answer to any of the seven questions. Changing the answers from negative to positive is a big challenge and sets the agenda for language testing in Austria (Sigott 2007:161).

The recognition of this state of affairs led to reactions at all levels in the educational system. Educational standards testing was introduced at the end of primary school and at the end of the lower secondary level. At the end of upper secondary level, a new centralised Matura exam was implemented, and attempts were made to standardise the testing of achievement at university exit level in the language subjects. These developments are described in the contributions to Part I.

2. Exam reform projects: Educational Standards, Matura, ELTT Initiative

The first decade of the 21st century saw a growing interest in descriptions of learning goals in the language subjects. This was helped by the availability of the level descriptors of the CEFR, which were seen as welcome formulations of teaching targets for foreign languages, particularly English. Educationists, politicians and practitioners started talking about competence-based teaching. Eventually, policy makers propagated the idea of measuring attainment of these targets for the purpose of system monitoring (Lucyshyn 2004; 2007). This led to the introduction of the National Educational Standards (Bildungsstandards) in the core subjects German as a first language, Mathematics, and English as a foreign language. Once teaching targets had been formulated, it was decided to monitor attainment of these targets by means of standardised testing systems and to provide feedback to stakeholders, both at individual and at institutional levels. Accordingly, educational standards tests were developed to test achievement in German as a first language and in Mathematics at the end of primary school, and for German as a first language, Mathematics and English at the end of the lower secondary level. The tests for German as a first language cover reading and writing at both the end of primary school and at the end of the lower secondary level while English at lower secondary level is tested in the four skills of listening, speaking, reading and writing. A new institution, BIFIE (Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens) was created and put in charge of exam development with professional support from the universities of Vienna and Klagenfurt.

The contributions in Part I are arranged according to the level in the educational system that they refer to. The competence models for L1 German in the National Educational Standards and their operationalisation as test tasks and items are described by Eibl, Schwaiger and Breit. This is followed by a report by Luger-Bazinger, Itzlinger-Bruneforth and Schreiner on the development of a procedure for providing criterion-related feedback to individuals and to institutions,

taking L1 German reading as an example. For English, the test constructs and their operationalisation are described by Mewald, Allan and Kulmhofer (Listening), Siller and Kulmhofer (Reading), Mewald (Speaking) as well as Kulmhofer and Siller (Writing).

BIFIE was also put in charge of implementing a new high-stakes Matura exam for the end of the upper secondary level. This meant creating standardised tests for all the subjects, including Mathematics, for which written examinations were prescribed. Thus, for the language subjects, standardised exams had to be developed for testing achievement in the language of instruction, predominantly German but also Hungarian, Croatian and Slovene, for the so-called modern languages, namely English, French, Italian and Spanish, and for the classical languages Ancient Greek and Latin. This monumental task was undertaken jointly by BIFIE and the universities of Innsbruck, Klagenfurt, Vienna and external consultants. Exam development for the modern languages was spearheaded by a project started by the University of Innsbruck with the support by external consultants from Great Britain. This complex and demanding task involved creating tests for Listening, Reading, Writing and Language in Use for English, French, Italian and Spanish at different levels for candidates with eight, six and four years of instruction in different languages, respectively. All these exams were linked to the CEFR. The development of the exams for German, Hungarian, Croatian and Slovene as first languages or languages of instruction as well as the development of the exams for the classical languages were supported by expertise from the universities of Klagenfurt, Vienna and external consultants from Germany and Switzerland. Unlike in the case of the modern languages, these exams could not be linked to an external scale like the CEFR because such scales are not available. Struger describes the theoretical bases underlying the written exam for German as a language of instruction and their operationalisation. The field testing of a large number of tasks developed according to these principles is described by Cesnik and Sigott. Spöttl, Eberharter, Holzknecht, Kremmel and Zehentner provide an account of the process and challenges of implementing competence-based assessment for the modern languages English, French, Italian and Spanish while Weiler and Frötscher describe exam maintenance and quality control procedures for the exam now that it has been implemented. The way the new exam philosophy has been adopted in the classical languages Latin and Greek is described by F. Lošek, Glatz, Niedermayr and Weyrich-Zak, and selected aspects of its implementation, including washback effects, are discussed by Sörös, Seitz, Glas, Kuchling, Oswald and A. Lošek.

At university exit level, a group of university language teachers from the universities of Graz, Innsbruck, Klagenfurt, Salzburg and Vienna constituted the English Language Testing and Teaching Initiative (ELTT), led by the Klagenfurt Language Testing Centre. This initiative aimed at formulating level descriptors for assessing achievement in English writing and speaking. These were intended for assessing students majoring in English language and students in the English teacher training programmes at the end of their university language training. Analytic rating scales were developed on the basis of the CEFR in a combination of theory-based and empirical approaches and benchmarked performances were made available. Unlike in the case of the educational standards and Matura, the use of these scales is not compulsory, but it is up to the individual university department to decide whether or not to use them. These developments are described by Berger and Heaney.

Finally, the testing of German as a Foreign or Second language at all six levels of the CEFR in the institutional framework of the Österreichisches Sprachdiplom (ÖSD) is described by Glaboniat and Peresich.

3. The research scenario

The existence of national tests and examination systems with candidatures between 25,000 and 80,000 has given rise to a considerable variety of research. This is remarkable given the short time that has elapsed since the examinations entered the operational phase. Research has been carried out on tests or examinations at all the levels of the educational system ranging from primary to university level. This research is presented in Part II. The contributions in Part II are also arranged according to the level in the educational system that they refer to. Within each level, the research studies are grouped according to an adapted version of Weir's (2005) validity framework shown in Figure 1. This identifies six aspects of validity, which interact with each other and also overlap. In a tentative spirit, the contributions have been entered into the individual boxes according to the editor's interpretation of the original framework as well as the editor's perception of the main focus of each study. The allocation of studies to the individual aspects of validity is not intended to be prescriptive in any sense but is meant to aid the reader's orientation throughout the volume.

Fig. 1: Mapping of research foci against Weir's validity framework (based on Weir, 2005, pp. 44–47)

Context validity	**Theory-based validity**
Siller & Kipman Pibal, Sigott & Cesnik Maurer Zehentner	Siller & Kipman Pibal, Sigott & Cesnik Holzknecht Eberharter Maurer Heaney
Scoring validity & Standardsetting	**Rating validity & Benchmarking**
Sigott & Cesnik Platzer	Freunberger, Breit & Illetschko Cesnik Eberharter, Holzknecht & Kremmel Berger
Consequential validity	**Criterion-related validity**
Mewald (Part II) Frötscher, Themessl-Huber & Weiler Frötscher Platzer	Sigott & Cesnik Frötscher, Themessl-Huber & Weiler Platzer

Context validity refers to all facets of the test context from physical features like background noise or lighting to expectations held by stakeholders, which includes criteria for correctness. **Theory-based validity** is a cover term for the aspects of the test construct that are accessed by the test, particularly the cognitive processes that are triggered by the test tasks. **Scoring validity and standardsetting** relates to the process of translating test performance into numerical format for item-based tests while **rating validity and benchmarking** refers to the analogous process for tests of the productive skills. **Criterion-related validity** encompasses correlational approaches as well as attempts at relating test scores to well established published performance level descriptors, as is typically the case in attempts to link tests to the CEFR. **Consequential validity** refers to the effects that tests have for stakeholders. This includes effects on teaching, learning, and formative as well as summative assessment practices in the classroom.

In the area of the educational standards, context validity, theory-based validity, scoring validity and standardsetting, rating validity and benchmarking, criterion-related validity as well as consequential validity have been addressed to different extents. **Context validity** and **theory-based validity** is addressed in a study by Siller and Kipman, who examine contextual and cognitive factors that account for differences in item difficulty in the E8 reading test. Pibal, Sigott and Cesnik have studied the role that language errors play in the rating of writing performances

in the E8 writing baseline test, thus shedding light on criteria for correctness that are consciously or subconsciously applied by the raters. Problems **of scoring validity and standardsetting** are addressed in a study by Sigott and Cesnik, who, after describing approaches to standardsetting for the receptive skills in English, focus particularly on the discrepancies between empirical item difficulty and the difficulty estimates made by a standardsetting panel and ways of dealing with these discrepancies without asking panellists to negotiate agreement. Freunberger, Breit and Illetschko report on a study in the area of **rating validity and benchmarking**. They describe the rater training for the writing part of the German test at the end of primary school and document changes in rater behaviour from the training phase to operational testing with particular focus on rater severity and rater agreement. In particular, they identify texts with regard to which the raters tend to disagree and suggest research into features which are characteristic of such controversial texts. Cesnik focuses on an important methodological aspect of Multifacet Rasch Measurement, namely the algorithm for computing so-called fair measures, which are a frequently used statistic reported in studies of rating validity and benchmarking. **Criterion-related validity** is also addressed by Luger-Bazinger, Itzlinger-Bruneforth and Schreiner in Part I, who describe procedures for linking test scores for the receptive skills to level descriptors of the competence model for L1 German. Sigott and Cesnik (see above), describe and evaluate procedures for linking test scores for the receptive skills in English to the CEFR. Eventually, **consequential validity** is addressed by Mewald, who attempts to trace washback effects of the English standards tests at lower secondary level with regard to teaching, assessment practices as well as students' attitudes and motivation to learn.

The research focusing on Matura addresses context validity, theory-based validity, rating validity and benchmarking, criterion-related validity as well as consequential validity. In the area of **context validity**, Maurer investigates linguistic features characteristic of blogs, a new text type among the text types required in the writing component. An important aim of this research is to contribute towards concretising criteria for correctness that are applied in the rating process for this genre. Zehentner studies the effect of partner choice in the oral pair interview, which was introduced in 2012. The focus of interest is on whether students are able to choose partners in a way that is beneficial to their performance and to the subsequent rating. His study is exceptional in that it concerns the oral component of Matura, which to date has not been standardised to the extent the written components are. Three studies have a focus on **theory-based validity**. Holzknecht investigates the reading strategies that are triggered in test takers of the reading

component of Matura by means of think-aloud protocols and compares them to experts' judgments as to what individual items are testing. In an analogous study for the language in context component, Eberharter probes into the thought processes activated in test takers when processing the multiple-choice items of this component and relates them to expert judgments of item focus. Maurer's study (see above) also touches upon theory-based validity by identifying aspects of competence that are employed in the process of writing blogs as a task type in the writing component. The study by Eberharter, Holzknecht and Kremmel is highly relevant in the context of **rating validity and benchmarking.** They report on the effectiveness of a training programme for raters who double-rate writing performances for which this is deemed useful or for which this is required by the exam regulations. Their focus is on identifying facets of the training programme that are crucial in maximising rater consistency. Aspects of **criterion-related validity** are addressed by Frötscher, Themessl-Huber and Weiler. They report on the results of the 2016 Matura and focus on the effect of gender, school type, region, level of urbanity and years of instruction for French, Italian and Spanish, thus investigating potential test bias. By comparing the CEFR levels reached by the candidature to the stipulations of the individual curricula, they also touch upon washback, thus contributing evidence of **consequential validity**. Frötscher's study addresses the issue of washback more directly by investigating in detail the effect of the new Matura exam on the testing of reading in summative classroom tests. In contrast to Mewald (see above), who investigates the washback of the national standards tests, Frötscher finds clear evidence of washback of the new Matura exam. Aspects of washback in the classical languages are also described in the contribution by Sörös et al. in Part I, who report on developments in textbooks, assessment and in-service teacher training as a consequence of the new Matura exam.

The language testing research on assessment tools at university exit level is focused on rating validity and benchmarking, theory-based validity and on criterion-related validity. In the area of **rating validity and benchmarking** Berger reports on a series of studies aimed at providing validity evidence with regard to the ELTT speaking scales for presentations and interaction. He investigates the proficiency continuum posited by the scales using a multi-method approach that involves descriptor sorting, descriptor calibration and descriptor-performance matching. On the basis of the results, he suggests two modified versions of the ELTT scales and, more generally, detail to be added to the descriptors for the CEFR levels C1 and C2. Heaney's study deals with expeditious reading and refers primarily to **theory-based validity**. She suggests a definition for the construct of speed reading, a much-neglected aspect in the testing of reading. Based on a

review of the relevant literature, she suggests ways of operationalising expeditious reading strategies, describes a prototype version of a test of expeditious reading, and identifies directions for further research into operationalising expeditious reading strategies. Platzer's contribution has a focus in **criterion-related validity** but also deals with aspects of scoring validity and standardsetting, and of consequential validity. He examines issues of validity in vocabulary testing at tertiary level using criterion-referenced rather than norm-referenced methods.

The contributions to this volume show that not only has there been considerable progress in test development but that research activities are gaining momentum. In fact, a variety of research directions are being pursued, focusing on testing at all levels in the educational system. Further research is needed in the areas of washback, rating scale validation, cognitive strategy use and genre-specific features of the text types required in writing. A central area that will need more attention in the future is research into causes of differences in task or item difficulty in all languages tested. Studies along this line will need to be widened in scope to include a wider range of contextual and cognitive features, and to also include the productive skills, particularly writing. Here the focus will need to be on text features that predispose raters to award particular scale scores to writing performances. With reference to the test constructs of interest, the features studied could be categorised as those which are, and those which are not, supposed to affect task or item difficulty. The empirical data will then make it possible to determine the extent to which the test developers' intentions have been realised in the test tasks or items. While this kind of research is of obvious relevance for test validation, it is equally relevant for standardsetting. The better the members of standardsetting panels understand and perceive factors which determine item difficulty, the more agreement there will be among panellists but also, and more importantly, between the panellists' difficulty estimates and the empirical task or item difficulties. Research along these lines also has obvious relevance for teaching. The more is known about features affecting task and item difficulty, the better we understand the nature of the constructs, and, importantly, the nature of the competences which constitute the teaching goals. In fact, this kind of research could serve as the hub for future studies as it connects with most validation research in a more or less direct way.

References

Lucyshyn, J. (2004). Implementierung von Bildungsstandards in Österreich. *Erziehung und Unterricht* 7-8 (154. Jg.), S. 613–617.

Lucyshyn, J. (2007). Bildungsstandards in Österreich – Entwicklung und Implementierung: Pilotphase 2 (2004–2007). *Erziehung und Unterricht* 7-8 (157. Jg.), S. 566–587.

Sigott, G. (2007). A case for language testing. In M. Reitbauer, N. Campbell, S. Mercer & R. Vaupetitsch (Eds.), *Contexts of English in Use* (pp. 159–166). Vienna: Braumüller.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.

Sandra Eibl, Silke Schwaiger & Simone Breit

Bildungsstandardüberprüfung Deutsch: Kompetenzmodelle und Aufgabenentwicklung

Educational standards and standards testing in Austria aim to further improve educational quality in schools and to enhance teaching. These standards define which competencies pupils should have at the end of the 4th and the 8th school year. On behalf of the Austrian Federal Ministry of Education the Federal Institute for Educational Research, Innovation and Development of the Austrian School Sector (BIFIE) is in charge of further developing, implementing and the monitoring educational standards in Austria. This section focuses on German standards testing in Austria. It illustrates the principles of standards testing from the development of competence models, which are a crucial foundation of educational standards besides curricula, through the production of test items within various fields of competencies and their psychometric analyses to the use of these items in testing.

Claudia Luger-Bazinger, Ursula Itzlinger-Bruneforth & Claudia Schreiner

Die Umsetzung der kriterialen Rückmeldung in der Bildungsstandardüberprüfung am Beispiel Lesen in Deutsch für die vierte und achte Schulstufe

Criterion-referenced feedback is an important part of large-scale assessment, as it establishes a relationship between the test result and specific competencies. Developing a criterion-referenced feedback procedure is a process with multiple steps. After establishing a competence structure model that defines the competence and also sets the frame for the test and the writing and piloting of test items, a performance level description (or competence level model) is needed. It is used for setting the relevant standards and entails a specific description of the competences required for each of the competence levels. This is a prerequisite for a standardsetting procedure, in which scores are defined by an expert group. This complex process also needs evaluation to ensure valid criterion-referenced feedback. This contribution illustrates the way to criterion-referenced feedback, taking the reading assessment in German in the Austrian National Educational Standards (Bildungsstandardüberprüfung) as an example.

Claudia Mewald, Dave Allan & Andrea Kulmhofer

The development of the Austrian Educational Standards Test for English Listening at grade 8

Dieser Beitrag beschreibt die Genese des Hörverstehenstests für Englisch in der achten Schulstufe im Rahmen der Österreichischen Nationalen Bildungsstandards. Zunächst werden die theoretischen Grundkonzepte beschrieben, aus denen das Testkonstrukt hervorgegangen ist. Sodann werden die Testspezifikationen präsentiert, an deren Formulierung das Klagenfurter Language Testing Centre, das Norwich Institute for Language Education und die Universität Lancaster in chronologischer Reihenfolge beteiligt waren. Die stufenweise Entwicklung des Testkonstrukts, der Prozess der Itementwicklung und die Verfahren zur Anbindung des Tests an den GERS werden beschrieben. Schließlich wird die Art, in der die Rückmeldung der Resultate erfolgt, präsentiert.

Klaus Siller & Andrea Kulmhofer

The development of the Austrian Educational Standards Test for English Reading at grade 8

Im Jahr 2009 wurden die Bildungsstandards in den Fächern Mathematik, Deutsch und Englisch gesetzlich verankert und die regelmäßige Überprüfung der in den Bildungsstandards beschriebenen Lernziele gesetzlich festgelegt. Dieser Beitrag beschreibt die Entwicklung des Tests für Leseverständnis in Englisch als Fremdsprache auf der achten Schulstufe. Die unterschiedlichen Zugänge und Entwicklungen im Hinblick auf die Überprüfung der Fertigkeit Lesen werden dokumentiert, wobei ein Bogen von der Pilotphase (2005-2008) bis hin zur zweiten österreichweiten Überprüfung (2019) gespannt wird. Die Genese des Testkonstrukts, dessen Operationalisierung, Testdurchführung sowie das Konzept der Ergebnisrückmeldung werden beschrieben.

Claudia Mewald

The development of the Austrian Educational Standards Test for English Speaking at grade 8

Dieser Beitrag beschreibt die Entwicklung und Durchführung des Tests für mündliche Produktion in Englisch auf der achten Schulstufe im Rahmen der Österreichischen Bildungsstandards. Zuerst wird der theoretische Hintergrund für die Entwicklung des Testkonstrukts sowie der Beurteilungsskala resümiert. Darauf folgt eine detaillierte Dokumentation des Testkonstrukts und der Beurteilungsskala, des Testablaufs sowie der Schulung von Aufgabenerstellerinnen, Assessoren und gesprächsführenden Personen (Interlocutors). Weiters werden der Prozess der Aufgabenentwicklung, das Verfahren zur Anbindung an den GERS sowie das Verfahren der Ergebnisrückmeldung beschrieben.

Andrea Kulmhofer & Klaus Siller

The development of the Austrian Educational Standards Test for English Writing at grade 8

Im Jahr 2009 wurden die Bildungsstandards in den Fächern Mathematik, Deutsch und Englisch gesetzlich verankert und die regelmäßige Überprüfung der in den Bildungsstandards beschriebenen Lernziele gesetzlich festgelegt. Dieser Beitrag beschreibt die Entwicklung des Tests für schriftliche Produktion in Englisch als Fremdsprache auf der achten Schulstufe. Die unterschiedlichen Zugänge und Entwicklungen im Hinblick auf die Überprüfung der Fertigkeit Schreiben werden dokumentiert, wobei ein Bogen von der Pilotphase (2005-2008) bis hin zur zweiten österreichweiten Überprüfung (2019) gespannt wird. Die Genese des Testkonstrukts, dessen Operationalisierung, Testdurchführung sowie das Konzept der Assessorenschulung und der Ergebnisrückmeldung werden beschrieben.

Jürgen Struger

Deutsch als Unterrichtssprache: Das Konzept der schriftlichen Reife- und Diplomprüfung

This contribution describes the theoretical basis and the design of the written matriculation exam for L1 German, as it has been implemented at Austrian schools since the academic year 2014/2015. After a discussion of the requirements that a competence-based exam format for assessing writing skills has to fulfil, the exam format is described. An attempt is made to derive the components of the exam format from the assumption that, ultimately, competences manifest themselves in 'writing acts' (Schreibhandlungen). Future challenges arising from the implementation of the exam format, particularly with regard to task development and the assessment of individual aspects of competence, are discussed.

Hermann Cesnik & Günther Sigott

Die Standardisierte Reife- und Diplomprüfung (SRDP) Deutsch als Unterrichtssprache: Methodenkonzept und Hauptergebnisse der Feldtestung

Writing tasks for high-stakes tests like the standardised matriculation exam for L1 German should be validated before operational use. If the validation procedure involves field testing, the aims are twofold. On the one hand, the tasks have to be tested for their usability in operational testing, and on the other, as a long-term perspecitve, they need to be incorporated into a calibrated item bank. To achieve this, a field testing design has to be developed which makes it possible to generate task statistics in the interplay of test taker, rater and rating scale. Although test takers are an integral component of the research design, their purpose is to enable insights into the effect of test taker characteristics, motivation and grade distributions to be expected in the operational phase of the exam. While the results of the field testing are of no consequence to the test takers, the statistics do have consequences for the task itself, for the raters and for the rating scale. The results may suggest modifications to the tasks, adjustments of the guidelines and the training of item writers, directions for rater training and eventually adaptations of the rating scale. This contribution describes the phases of the field testing carried out at selected schools in 2012 and 2013. The data base for the analysis are ratings on a five-level scale for the dimensions of CONTENT, TEXT STRUCTURE, STYLE and EXPRESSION, and ACCURACY. This data base is supplemented with structural data (e.g. school or school type, task characteristics (e.g. text type), data from a questionnaire (e.g. L1, bilingualism, gender) and Rasch-based estimates (e.g. measures, fit statistics, standard error)). Altogether 60 tasks from 2012 and 54 tasks from 2013 were field tested. In additon, four tasks were tested again in 2013 in order to enable linkage of the two field tests.

Carol Spöttl, Kathrin Eberharter, Franz Holzknecht, Benjamin Kremmel & Matthias Zehentner

Delivering reform in a high stakes context: from content-based assessment to communicative and competence-based assessment

2015 fand in Österreich die erste verpflichtende und flächendeckende Liveadministration der reformierten standardisierten Reifeprüfung statt. 15.000 SchülerInnen der Allgemeinbildenden Höheren Schulen bearbeiteten standardisierte, kompetenzorientierte, kommunikative Aufgabenstellungen in den Fertigkeiten Lesen, Hören, Sprache im Kontext und Schreiben. Der Prozess dieser Reform begann jedoch bereits mehr als 10 Jahre vor dieser ersten Administration mit der Einführung eines GERS-basierten nationalen Kurrikulums für die modernen Fremdsprachen. 2007 wurde einem Forschungsteam der Universität Innsbruck vom österreichischen Bundesministerium für Unterricht, Kunst und Kultur ein Pilotprojekt zur Erforschung und wissenschaftlichen Begleitung einer Einführung einer standardisierten Reifeprüfung zugesprochen. Dieses Kapitel beschreibt die Notwendigkeit und die Anfänge dieses Reformprojekts, das Österreich von inhaltsorientiertem zu kompetenzorientiertem kommunikativem Sprachtesten und von individuell von LehrerInnen erstellten internen Prüfungen zu einer teilzentralen standardisierten und professionell entwickelten Reifeprüfung geführt hat, die erhöhte Fairness für SchülerInnen garantiert. Es beschreibt Bewertungspraktiken vor der Reform und wie das Modell der neuen Matura entwickelt und bis zur Projektübergabe an das BIFIE etabliert wurde.

Theresa Weiler & Doris Frötscher

Ensuring sustainability and managing quality in producing the standardized matriculation examination in the foreign languages

Die österreichische standardisierte Reife- und Diplomprüfung (SRDP, Matura) in den lebenden Fremdsprachen entwickelte sich im Zuge einer Reform von einer lehrerzentrierten Prüfung zu einer standardisierten und zentral vorgegebenen Prüfung. Dieser Beitrag gibt Einblick in die Produktion und Qualitätssicherung der Prüfungsaufgaben. Die SRDP in den lebenden Fremdsprachen unterlag von der Anfangsphase 2007 bis heute großen Veränderungen. 2012 wurde die Aufgabenentwicklung von der Universität Innsbruck auf das Bundesinstitut BIFIE übertragen, das in weiterer Folge 2017 in das Bundesministerium für Bildung integriert wurde. Eine zweite wichtige Änderung war der Übergang der SRDP vom Schulversuch ins Regelschulwesen. Im Mai 2016 wurde die SRDP erstmals österreichweit an allen höheren Schulen durchgeführt. Diese Änderungen hatten Auswirkungen auf die Prozesse im Testentwicklungszyklus. Der Umgang mit organisatorischen Veränderungen und politischen Herausforderungen wird in diesem Beitrag beschrieben. Des Weiteren liefert der Beitrag einen Überblick über die Prüfung und stellt die Schritte des Testentwicklungszyklus im Detail dar. Abschließend werden Einblicke in zukünftige Entwicklungen und Herausforderungen gegeben.

Fritz Lošek, Peter Glatz, Hermann Niedermayr & Irmtraud Weyrich-Zak

Entwicklung von Testsystemen in den klassischen Sprachen

für Wilma Widhalm-Kupferschmidt in memoriam

Teaching and assessment in the classical languages Latin and Greek have seen a radical paradigm shift in the last decade. This contribution starts with a survey of these changes and innovations. Then the basic structure and the most important principles of assessment (competence orientation, achievement-oriented rather than deficit-oriented marking, separation of translation from interpretation) are presented with a focus on the new matriculation exam (Reifeprüfung). This is followed by a report on comprehensive pilot tests, which preceded the new matriculation exam. The results of these pilot tests had important consequences for the final format of the exam but also, at the same time, produced washback effects on teaching. Eventually the implementation of the new test systems in the context of the written matriculation exam is explained.

Michael Sörös, Martin Seitz, Renate Glas, Walter Kuchling, Renate Oswald & Andrea Lošek

Aspekte der Implementierung der neuen Testsysteme in den klassischen Sprachen

für Wilma Widhalm-Kupferschmidt in memoriam

This contribution starts by discussing the relationship between the new assessment philosophy and the principle of competence orientation on the one hand, and the legal framework governing the Austrian school system on the other. It then describes the washback that the paradigm shift in assessment practice has produced in teaching materials, textbooks and in the attractiveness of the classical languages as school subjects. This is followed by the presentation of an innovative electronic marking and feedback support system which was developed in the wake of the new assessment principles and is now in use for classroom tests as well as for the written part of the matriculation exam in the classical languages. Another facet described is the formative assessment system and accompanying materials, which parallel the educational standards tests in English and German with the aim of directing learners at early stages towards the requirements in the final years of teaching and those of the written matriculation exam. The contribution ends with a look at the effect that the new assessment philosophy has had on in-service teacher training for the classical languages.

Armin Berger & Helen Heaney

Developing rating instruments for the assessment of Academic Writing and Speaking at Austrian university English departments

Dieser Beitrag beschreibt ein im österreichischen Hochschulkontext bis dato beispielloses Projekt zur Entwicklung analytischer Skalen für die Beurteilung von Sprachkompetenz im tertiären Bereich. Im Bestreben, die gängige Testpraxis an den österreichischen Anglistikinstituten zu professionalisieren, erstellte die English-Language-Teaching-and-Testing-Initiative (ELTT) drei Beurteilungsskalen für wissenschaftliches Schreiben, Präsentationen und Interaktionen auf C1 und C2 gemäß Gemeinsamem Europäischen

Referenzrahmen für Sprachen. Der Beitrag skizziert zunächst den Kontext des Projekts. Im Anschluss folgt eine detaillierte Beschreibung des Entwicklungsprozesses, der sich über eine Reihe von Workshops zwischen Juli 2006 und Juli 2010 erstreckte und theoriebasierte, intuitive sowie datengestützte Methoden der Skalenentwicklung kombinierte. Zunächst wurden sogenannte Ankerdeskriptoren aus den Referenzniveaus C1 und C2 extrahiert. Diese wurden dann mithilfe von konkreten Performanzen erweitert und durch eine ausdifferenzierte Zwischenstufe ergänzt. In einem iterativen Überarbeitungsprozess wurden die Formulierungen erprobt und schrittweise geschärft. Der Fokus des Beitrags wechselt anschließend auf die entstandenen Produkte, welche neben den Skalen noch Benchmark- Performanzen samt schriftlichen Begründungen für die jeweiligen Niveauzuteilungen umfassen. Abschließend folgen eine kritische Betrachtung des Projekts sowie zukünftige Handlungsperspektiven.

Manuela Glaboniat & Carmen Peresich

Österreichisches Sprachdiplom Deutsch

The ÖSD (Österreichisches Sprachdiplom Deutsch) is a state-approved and internationally recognized examination and assessment system for „German as a Foreign Language / German as a Second Language", which offers children, adolescents and adults the opportunity to have their language proficiency certified on all six levels of the Common European Framework of Reference for Languages (CEFR). All ÖSD exams are standardized proficiency as well as high-stakes tests, which serve as a reliable basis for a large number of decisions, e. g. in the fields of education, employment, or migration. After a brief introductory section covering not only the history of ÖSD but also the range of exams currently offered, the aims and test-theoretical principles (e. g. communicative approach, real-life approach, pluricentric approach) underlying ÖSD exams are outlined. Following this, selected quality assurance measures of all relevant areas – from test development to test evaluation, such as peer-reviewing during the test construction process, training of examiners, or their constant requalification – are discussed and illustrated with the aid of examples.

Roman Freunberger, Simone Breit & Marcel Illetschko

Beurteilerübereinstimmung und schwer zu beurteilende Texte im Vergleich Training und Bildungsstandardüberprüfung in Deutsch auf der vierten Schulstufe

The Austrian standards-based proficiency test for 4th graders in German language comprises, amongst tests of other sub-competencies, an assessment of students' written performances. The short texts are judged by expert raters who underwent highly standardized training which involved benchmark texts. As raters' judgements are prone to rater bias or rater error such as leniency/severity or inaccuracy, monitoring of rater behavior during training is mandatory to ensure reliable test scores in the field test. Previous research has tended to focus on pure rater effects. Here, we would like to establish a link to other findings that showed that rater inaccuracy seems not to be constant over texts. In fact, some texts are harder to rate than others. Following this line of reasoning we apply an accuracy index to our training data which can identify hard-to-rate texts. Also, we tackle the question of the consistency of rater agreement between training and field test. Our findings provide evidence that rating quality slightly decreases over time and that some texts give rise to difficulties during the rating process which might be due to text features, which future studies should focus on. The findings are discussed with respect to their relevance for improving rater training.

Günther Sigott & Hermann Cesnik

On the discrepancy between rater-perceived and Empirical Item Difficulty. Standardsetting for the receptive skills in the Austrian Educational Standards Tests for English

Dieser Beitrag thematisiert die Diskrepanz zwischen empirischer und durch Assessoren eingeschätzter Itemschwierigkeit am Beispiel der Kompetenzstufensetzung im GERS für die rezeptiven Fertigkeiten in der Standardstestung für Englisch in der achten Schulstufe. Es gibt verschiedene Ansätze zur Setzung von Kompetenzstufen (Standardsetting), die jedoch alle zwei Arten potentieller Unschärfen implizieren. Die eine resultiert aus Unterschieden in der Strenge der Assessoren und führt dazu, dass Items als unterschiedlich schwierig eingeschätzt werden. Die andere Unschärfe resultiert aus der Diskrepanz zwischen tatsächlicher (empirischer) und durch die Assessorinnen eingeschätzter Itemschwierigkeit. Der vorliegende Beitrag beschreibt zwei Ansätze, die es ermöglichen, trotz vorhandener Diskrepanz zwischen wahrgenommener und empirischer Itemschwierigkeit zu nachvollziehbaren Kompetenzstufensetzungen zu gelangen und Testteilnehmer in ein GERS Niveau für Leseverstehen und Hörverstehen einzuordnen. Dafür müssen Thresholds für die Einstufung der Items auf der GERS-Skala definiert werden. Das Verfahren zur Kompetenzstufensetzung beinhaltet drei Schritte. Zunächst werden mittels einer modifizierten Angoff-Prozedur auf der Basis der Einschätzung durch die Assessoren Raw Difficulty Estimates bestimmt. Sodann werden aus Testadministrationen Empirical Item Difficulties ermittelt. Schließlich werden unter Anwendung von Prinzipien der probabilistischen Testtheorie zwei Verfahren, der Fuzzy-Intervall Ansatz und der Regressionsansatz, beschrieben, die es ermöglichen, angesichts der Diskrepanz zwischen geschätzter und empirischer Itemschwierigkeit Kompetenzstufen zu setzen.

Florian Pibal, Günther Sigott & Hermann Cesnik

The role of error in assessing English Writing in the Austrian Educational Standards Baseline Test

Gegenstand dieser Untersuchung ist die Rolle von Fehlern in der Beurteilung von Schreibperformanzen in der Baseline-Studie der österreichischen Bildungsstandardstestung auf der achten Schulstufe. Die Rolle, die Fehler bei der Beurteilung spielen, ist sowohl für die Testvalidierung von Bedeutung als auch für die Praxis des Unterrichts. Um dieser Frage nachzugehen, wurde eine Fehleranalyse von 100 Schreibperformanzen durchgeführt. Fehler wurden nach linguistischer Beschreibungsebene (Grammatik, Lexis, Orthographie und Interpunktion, Diskurs) kategorisiert. Die Beurteilungen der Performanzen auf der analytischen Skala durch ausgebildete Assessoren wurden als abhängige Variablen in Korrelationsanalysen verwendet. Es bestehen signifikante starke bis mäßige Korrelationen zwischen den Fehlerkategorien und der aggregierten Gesamtbeurteilung sowie den einzelnen Beurteilungsdimensionen. Weitere Details der Zusammenhänge werden mittels Decision Tree-Analysen untersucht. Grammatikfehler erweisen sich als stärkste Determinante der Beurteilungen auf allen vier Dimensionen während lexikalische Fehler eine sehr geringe Rolle spielen. Fehler auf der Diskursebene sind Determinanten der Beurteilungen auf den Dimensionen für COHERENCE & COHESION sowie für TASK ACHIEVEMENT, doch scheint es für die Assessoren schwierig zu sein, die beiden Beurteilungsdimensionen klar voneinander getrennt zu halten. Orthographie- und Interpunktionsfehler determinieren die Beurteilungen auf allen vier Dimensionen, jedoch durchwegs in einer untergeordneten Rolle.

Claudia Mewald

The washback of standardised testing in the subject English at lower secondary level

Diese Studie beruht auf unterschiedlichen Konzepten zum „Washback“ (Alderson & Wall, 1993; Shohamy, 1993; Watanabe, 2004; 1997). Sie versucht, die Auswirkungen der standardisierten Überprüfung der Fremdsprachenkompetenzen auf der achten Schulstufe in Hinblick auf generelle und spezifische Veränderungen im Unterrichten und Lernen sowie in der Beurteilung zu beschreiben, deren Intensität einzuschätzen und die Erfüllung curricularer Vorgaben darzustellen. Im Zentrum des Interesses stehen die Auswirkungen der standardisierten Bildungsstandardtestungen in Englisch auf den Unterricht, das Lernerverhalten und die Leistungsmessung. Es zeigt sich, dass der Unterricht bereits Innovationen abbildet, die in Konstrukten und Testspezifikationen ausreichend dargestellt sind. Lehrkräfte implementieren sowohl Testdesigns als auch Beurteilungskonzepte vorwiegend im produktiven Bereich. Der Unterricht in den rezeptiven Fertigkeiten sollte durch Veröffentlichung der Test-Konstrukte und Schulung der Lehrkräfte konkrete Impulse zur Veränderung erhalten. Mit stärkerem Fokus auf das Lernen wäre dann auch mehr positiver Washback auf das Lernerverhalten zu erwarten.

Klaus Siller & Ulrike Kipman

Cognitive processes as predictors of item difficulty in the Austrian Educational Standards Baseline Test for English Reading at grade 8

2009 wurden die österreichischen Bildungsstandards für das Unterrichtsfach Englisch, erste lebende Fremdsprache, auf der 8. Schulstufe (E8 Standards) gesetzlich verankert und in einer Ausgangsmessung (E8 Baseline Test) zum ersten Mal überprüft. Der vorliegende Beitrag beschäftigt sich mit der Konstruktvalidität des E8 Baseline Lesetests aus dem Jahr 2009. Ausgehend von einer allgemeinen Beschreibung von Validierung und kognitiver Leseprozesse werden im Anschluss aus einer Vielzahl von kognitiven Prozessen, die bei der Bearbeitung der 120 untersuchten E8-Leseitems eine Rolle spielen dürften, Prädiktoren der empirischen Aufgabenschwierigkeit im statistischen Sinn identifiziert. Items und deren empirische Schwierigkeitsmaße wurden vom Bundesinstitut BIFIE zur Verfügung gestellt. Es zeigt sich ein signifikanter Zusammenhang zwischen Itemschwierigkeit und den kognitiven Prozessen metacognitive activities und processing operations. Abschließend werden Implikationen für die Aufgabenerstellung und den Leseunterricht in der Fremdsprache diskutiert.

Doris Frötscher, Michael Themessl-Huber & Theresa Weiler

Analyzing the 2016 standardized matriculation exam results in the foreign languages: post-test analysis

Die standardisierte Reife- und Diplomprüfung (SRDP) in den Lebenden Fremdsprachen Englisch, Französisch, Italienisch und Spanisch gilt seit 2016 in allen höheren Schulen österreichweit. Nach erfolgter Durchführung der schriftlichen SRDP ist eine Analyse der Klausurergebnisse unabdingbar für eine nachhaltige Qualitätssicherung (EALTA, 2006). Dieser Beitrag präsentiert Ergebnisse der Post-Test Analyse 2016 für Englisch AHS (Allgemeinbildende Höhere Schulen) und BHS (Berufsbildende Höhere Schulen) und

Französisch AHS. Die dazu analysierten Daten auf Kandidatenebene (anonymisiert) beinhalten u.a. das Geschlecht, die erreichten Punkte pro Aufgabe bzw. Kriterium, und die Gesamtnote. Unter Verwendung der Statistiksoftware R wurden deskriptive Statistiken berechnet und Analysen hinsichtlich Geschlecht, Schultyp und Lernjahre durchgeführt. Die Post-Test Analyse spielt eine wichtige Rolle im SRDP-Qualitätszyklus und liefert die Grundlage für eine Reihe von Schlussfolgerungen. Die Ergebnisse aus 2016 bestätigen, dass das vom Lehrplan vorgegebene Zielniveau für die österreichischen Schüler/innen erreichbar ist. Weiters tragen die Ergebnisse zu Monitoring und Validierung der Prüfung bei.

Doris Frötscher

Matura washback on the classroom testing of Reading in English

Diese Studie untersuchte Auswirkungen der neuen österreichischen Standardisierten Reife- und Diplomprüfung (SRDP, Matura) auf die Gestaltung von Leseaufgaben in Schularbeiten. Auswirkungen von Tests werden immer häufiger als Bestandteil von Testvalidität betrachtet; dennoch war bislang wenig bekannt über die Auswirkungen standardisierter Tests auf Prüfungen, die von LehrerInnen lokal auf Klassenebene eingesetzt werden. Die in diesem Beitrag präsentierte Studie war Teil eines Doktoratsprojektes, das eine Bearbeitung dieser Forschungslücke zum Ziel hatte. Hierzu wurden 173 Schularbeiten-Leseaufgaben aus den Schuljahren 2001-2013 analysiert und die Aufgabencharakteristika der beiden Zeiträume vor bzw. nach der Einführung der SRDP 2008 verglichen. Das Forschungsdesign zeichnet sich durch die Verwendung eines speziell entwickelten Aufgabenanalyseinstruments aus, das nicht von den Darstellungen der Teilnehmer/innen abhängig ist. Die Ergebnisse zeigten eine beträchtliche Änderung der verwendeten Testmethoden sowie eine extensive Verwendung von SRDP-Aufgaben im Zeitraum nach 2008. Der Beitrag diskutiert, inwiefern diese Änderungen als positiv oder negativ anzusehen sind, und schließt mit Empfehlungen für die weitere Implementierung der SRDP und mit Ideen für weitere Forschung.

Michael Maurer

How do test takers write a blog? A corpus-based study of linguistic features in performances based on the task type Blog

Blogs wurden als schriftliche Textsorte der österreichischen Reife- und Diplomprüfung (Matura) implementiert. Es fehlt allerdings noch an Forschungsergebnissen, die beschreiben, welche sprachlichen Merkmale Blogs aufweisen, vor allem wie Lernende diese Textsorte behandeln. Dies führt dazu, dass sich Sprachlehrende und -tester nicht sicher sind, was sie von Lernenden erwarten können. In diesem Beitrag wird eine Studie präsentiert, die untersucht, welche sprachlichen Merkmale TestkandidatInnen in ihren Blogs verwenden und ob in ihren Blogs unterschiedliche sprachliche Merkmale im Vergleich mit Essays verwendet werden. Es werden Ergebnisse einer korpusbasierten Studie von 198 Lernerperformanzen präsentiert. Sie identifiziert sprachliche Merkmale, die ausschließlich in Blogs von TestkandidatInnen auf B2-Niveau zu finden sind und sich von Essays von derselben Population unterscheiden. Ein Vergleich der Blog- und Essay-Subkorpora mit der Academic Word List (Gardner & Davies, 2013) und der PHRASE List (Martinez & Schmitt, 2012) zeigt, dass spezielle Ausdrücke in beiden Textsorten vorkommen, aber in unterschiedlichen Wortkombinationen verwendet werden.

Franz Holzknecht

Reading strategies in the standardized Austrian matriculation examination for English

In dieser Studie wurde untersucht, welche Lesestrategien in der standardisierten österreichischen Reifeprüfung für Englisch (Matura) getestet werden. Zu diesem Zweck verbalisierten sieben TeilnehmerInnen ihre Gedanken, während sie den Lesetest der standardisierten Matura von 2011/12 bearbeiteten (concurrent think-aloud protocols). Außerdem bewerteten sechs ExpertInnen, welche Lesestrategien in den einzelnen Fragen getestet wurden. Die Ergebnisse der think-aloud Protokolle zeigen, dass zwar der Großteil der Lesestrategien, die laut Testentwicklern notwendig sind, um die Fragen zu beantworten, auch von KandidatInnen herangezogen, gewisse Strategien jedoch nicht ausreichend überprüft werden. Die Ergebnisse der ExpertInnenbewertungen machen darüber hinaus deutlich, dass sich die ExpertInnen nicht einig waren, welche Strategien überprüft werden. Die Studie gibt Aufschluss über die Validität des Tests in Bezug auf Lesestrategien sowie über die Eignung der verwendeten Untersuchungsmethoden zu Forschung in diesem Bereich.

Kathrin Eberharter

Validating multiple choice Language in Context items

Dieser Beitrag präsentiert eine qualitative Untersuchung der Konstruktvalidität von Multiple-Choice Grammatik- und Vokabelitems der neuen Standardisierten Reifeprüfung in Englisch. Die Studie stützt sich auf die Triangulation von ExpertenInneneinschätzungen (N=5) zur jeweiligen Ausrichtung der Items, Simultaneous Think-Aloud Protokollen (TAP) und Stimulated Recall Protokollen (SR) (Ericsson & Simon, 1993) von 13 ProbandInnen. Die beiden Protokolle wurden auf Mikro- und Makroebene segmentiert, kodiert und analysiert, und gewähren einen Einblick in die kognitiven Prozesse und Lösungsstrategien der ProbandInnen. Die Ergebnisse zeigen, dass die kognitiven Prozesse größtenteils durch die Ausrichtung der Items – grammatisch, lexikalisch oder lexiko-grammatisch – geprägt werden. Allerdings wurden auch Items identifiziert, bei denen die ProbandInnen unerwartete Denkprozesse aktivierten. Lösungsstrategien wie das Raten, die Aktivierung von Vorwissen, und das Ausschließen von Distraktoren wurden identifiziert und deren Rolle im Entscheidungsprozess untersucht. Abschließend werden neue methodologische Ansätze zur Verfolgung von Denkprozessen und weitere Desiderata der kognitiven Validierung von Sprachtests aufgezeigt.

Kathrin Eberharter, Franz Holzknecht & Benjamin Kremmel

Evaluating the effectiveness of a training program for double-raters

Die Bewertung der schriftlichen Leistungen in der neuen österreichischen Standardisierten Reife- und Diplomprüfung in den Fremdsprachen (Matura) obliegt trotz Standardisierung der Aufgabenstellung den individuellen Lehrpersonen. Da nur ein Teil der Lehrpersonen in der Anwendung des neuen GERS basierten Bewertungsrasters ausgebildet werden kann, wurde vom Bundesministerium in Kooperation mit der Universität Innsbruck und der Pädagogischen Hochschule Tirol ein Projekt zur Ausbildung einer Gruppe von ZweitbewerterInnen ins Leben gerufen. Insgesamt wurden in zwei Lehrgängen zwischen März 2011 und Februar 2013 32 Österreichische Lehrpersonen ausgebildet. Der vorliegende Beitrag beschreibt die Ausbildungsschritte im Detail und untersucht mittels einer Rasch-basierten Analyse den

Trainingseffekt der Ausbildung auf die TeilnehmerInnen und inwiefern sich die beiden Lehrgänge voneinander unterscheiden. Ziel dieser Arbeit ist es, den Erfolg und die Effektivität dieser Ausbildung zu evaluieren.

Matthias Zehentner

Examining the effect of partner choice in peer-to-peer assessment

Im Zuge der Reform der mündlichen österreichischen Reifeprüfung in den Fremdsprachen im Jahr 2012 wurde das davor vorherrschende Interviewformat durch ein zweigeteiltes Prüfungsformat ersetzt. Um SchülerInnen die Chance zu geben, ein breiteres Spektrum an Sprachfunktionen zu demonstrieren, ist seitdem zusätzlich zu einem Präsentationsteil eine Diskussionsaufgabe Teil der Matura. Ähnliche Formate finden sich bereits in einer Vielzahl von internationalen Tests, da sie unter anderem als eine authentischere Repräsentation von alltäglichen Konversationen als Interviewformate gesehen werden. Vorbehalte gegenüber Diskussionsaufgaben beruhen meist darin, dass ein Gespräch mit einem Peer ungewollte Variation einführen könnte. Insbesondere das sprachliche Niveau des Gegenübers wird als potentiell unfaire Schwankung gesehen. Eine gängige Art, mit dieser Problematik umzugehen ist, die SchülerInnen ihre/n Partner/in selbst auswählen zu lassen. Dieser Beitrag berichtet von einer Studie, deren Ziel es war zu untersuchen, ob SchülerInnen ihre/n PartnerIn zu ihrem Vorteil wählen können und inwieweit die Partnerwahl sich auf die Einstellungen der SchülerInnen auswirkt.

Hermann Cesnik

Annäherung der Transformationsfunktion von Logit Measures auf Fair Measures im Multifacetten-Raschmodell (MFRM) durch logistische Regression. Eine Simulationsstudie

The aim of this simulation study is to develop mathematical transformation functions to approximate the relation between MFRM-based Logit and Fair Measures. In MFRM this relation is represented as the Test Characteristic Curve. Two methodological approaches are distinguished, namely the general logistic regression approach and a modified logistic regression approach. The general logistic regression approach requires determination of the regression coefficient b_1 (slope in linear regression) and the constant term b_0 (intercept in linear regression). For the modified logistic regression approach the model parameters Slope and Location are derived directly from the empirical data for an interval of interest. This is possible without recourse to general logistic regression. Both approaches are applied to the two facets (main effects) of task and person. The quality of the approximation is assessed on the basis of these two main effects. This is done by comparing the approximated fair measures resulting from the general logistic approach and those from the modified logistic approach with the fair measures provided by Facets. Statistics for describing differences are developed and reported. For the data at hand, both approaches turn out to be equally suitable for modelling the Test Characteristic Curve.

Armin Berger

Rating scale validation for the assessment of spoken English at tertiary level

Während die Verwendung von Beurteilungsskalen für fremdsprachliche Leistungen an den philologischen Instituten der österreichischen Universitäten im Laufe der letzten Jahre zugenommen hat, sind die empirischen Untersuchungen zur Validität solcher Instrumente rar geblieben. Dieser Beitrag beschreibt indessen die mehrstufige Validierung zweier Skalen für die Beurteilung akademischer Präsentationen und Interaktionen. Die ersten drei Phasen zielten darauf ab, das in den Niveaubeschreibungen dargestellte Leistungskontinuum mithilfe von Korrelations- und Multifacetten-Rasch-Analysen zu untersuchen und die ursprünglichen Deskriptoren im Lichte der empirischen Ergebnisse zu modifizieren. Die letzte Phase hatte den Zweck, die Effektivität der revidierten Skalen zu prüfen. Hierfür wurde zum einen die psychometrische Qualität mittels Multifacetten- Rasch-Analyse quantitativ evaluiert, zum anderen wurde die Anwendbarkeit der Skalen anhand von retrospektiven Gruppeninterviews mit Beurteilenden qualitativ untersucht. Die Ergebnisse zeigen, dass die überarbeiteten Beurteilungsskalen die psychometrischen Qualitätskriterien weitgehend erfüllen. Ungereimtheiten ergaben sich vereinzelt in Bezug auf Niveaubeschreibungen am unteren Ende des Leistungskontinuums. Aus den vorliegenden Ergebnissen lassen sich nützliche Folgerungen für die Entwicklung und Validierung von Beurteilungsskalen allgemein ableiten.

Helen Heaney

Operationalizing expeditious reading strategies in tests of English as a foreign language

Es gibt Hinweise darauf, dass das effiziente, überfliegende Lesen größere Schwierigkeiten bereitet als das langsame, sorgfältige Lesen. Daher sollten Skimming und Scanning im (Fremd-) Sprach(en)unterricht gelehrt, geübt und geprüft werden. Zu diesem Zweck wurde ein umfassender Lesetest auf C1/C2-Niveau als Teil der Abschlussprüfung im Sprachkompetenzprogramm am Institut für Anglistik an der Alpen-Adria Universität Klagenfurt konzipiert. Dieser Beitrag beschreibt die von Weir und Khalifa (2008) geschilderten kognitiven Prozesse, die dem angewendeten Testkonstrukt zugrunde liegen sowie die Operationalisierung von überfliegendem Lesen durch innovative, computergestützte Prüfungsaufgaben mit vorgegebenen Lesegeschwindigkeiten. Anschließend werden einige der mit der Überprüfung von überfliegenden Lesestrategien verbundenen Schwierigkeiten andiskutiert, die im Rahmen der Entwicklung meiner Items zu Tage traten und die möglicherweise drei führende Testinstitutionen noch davon abhalten könnten, solche Strategien in ihre computerbasierten Englischprüfungen einzubauen. Ich schließe mit der Frage, ob eine Überprüfung der Effektivität von Skimming und Scanning, also eine Kombination von Geschwindigkeit und Genauigkeit, ein lohnenderes Ziel darstellen würde.

Hans Platzer

A criterion-referenced approach to the Vocabulary Levels Test: validating form B of the 2000 and 3000 word levels, and the Academic Wordlist

Paul Nations Vocabulary Levels Test (VLT) dient als diagnostisches Testinstrument zur Abschätzung der Wortschatzgröße in Englisch als L2. Trotz seiner weiten Verbreitung existieren nur wenige Validierungsstudien, die außerdem – bis auf eine Ausnahme – auf normorientierten Kriterien beruhen, obwohl der VLT als kriteriumsorientierter Test konzipiert wurde. Die gegenwärtige Studie geht daher von einem rein kriteriumsorientierten Ansatz aus. In diesem Zusammenhang belegen sowohl der B-Index als auch Item Phi eine ausreichende Trennschärfe der Mehrzahl der analysierten Itembündel, die auf der Basis von Phi Lambda und Kappa Quadrat eine zufriedenstellende Reliabilität der einzelnen Subtests von ≥0.80 ergeben. Weiters bestätigt eine Korrelationsanalyse die Testwirkungsvalidität (consequential validity) des VLT, da keine Benachteiligung von männlichen oder weiblichen KandidatInnen vorliegt, mit der bei Multiple-Choice-Verfahren in einzelnen Fällen zu rechnen ist. Eine Hauptkomponentenanalyse belegt, dass rezeptives lexikalisches Wissen das einzige zugrundeliegende Testkonstrukt dastellt, was die Kontruktvalidität des VLT unterstützt. Letztendlich bestätigt ein Vergleich des VLT mit zwei unabhängigen Sprachtests (X-Lex und electronic Quick Placement Test von Oxford University Press) dessen Kriteriumsvalidität. Der VLT hat daher ausreichende Reliabilität und Validität für seinen Einsatz als kriteriumsorientiertes diagnostisches Instrument im Englischunterricht.

Autorenverzeichnis

Dave Allan, NILE, dave@nile-elt.com
Armin Berger, Universität Wien, armin.berger@univie.ac.at
Simone Breit, PH Niederösterreich / BIFIE, simone.breit@ph-noe.ac.at / s.breit@bifie.at
Hermann Cesnik, Alpen-Adria-Universität Klagenfurt, hermann.cesnik@aau.at
Kathrin Eberharter, Universität Innsbruck, kathrin.eberharter@uibk.ac.at
Sandra Eibl, BIFIE Salzburg, s.eibl@bifie.at
Roman Freunberger, BIFIE Salzburg, r.freunberger@bifie.at
Doris Frötscher, BMBWF, Wien, doris.froetscher@bmbwf.gv.at
Manuela Glaboniat, Alpen-Adria-Universität Klagenfurt, manuela.glaboniat@aau.at
Renate Glas, Europagymnasium Klagenfurt, glas.agamemnon@utanet.at
Peter Glatz, Stiftsgymnasium Wilhering, peter.glatz@eduhi.at
Helen Heaney, Universität Wien, helen.heaney@univie.ac.at
Franz Holzknecht, Universität Innsbruck, franz.holzknecht@uibk.ac.at
Marcel Illetschko, BIFIE Salzburg, m.illetschko@bifie.at
Ursula Itzlinger-Bruneforth, BIFIE Salzburg, u.itzlinger@bifie.at
Ulrike Kipman, PH Salzburg Stefan Zweig, ulrike.kipman@phsalzburg.at
Benjamin Kremmel, Universität Innsbruck, benjamin.kremmel@uibk.ac.at
Walter Kuchling, BG/BRG Lerchenfeld, Klagenfurt, kuchling.agamemnon@utanet.at
Andrea Kulmhofer, BIFIE Salzburg, a.kulmhofer@bifie.at
Andrea Lošek, Pädagogische Hochschule Niederösterreich, andrea.losek@ph-noe.ac.at
Fritz Lošek, Landesschulrat für Niederösterreich, friedrich.losek@lsr-noe.gv.at
Claudia Luger-Bazinger, ÖZBF, claudia.luger-bazinger@oezbf.at
Michael Maurer, Universität Innsbruck, m.maurer@gemnova.at
Claudia Mewald, Pädagogische Hochschule Niederösterreich, claudia.mewald@ph-noe.ac.at
Hermann Niedermayr, Akademisches Gymnasium Innsbruck, hermann.niedermayr@tele2.at
Renate Oswald, BG Rein, renateoswald@aon.at
Carmen Peresich, ÖSD, carmen.peresich@osd.at
Florian Pibal, Magistrat der Stadt Villach, florian.pibal@villach.at
Hans Platzer, Fachhochschule Wiener Neustadt GmbH, hans.platzer@fhwn.ac.at
Claudia Schreiner, BIFIE Salzburg, c.schreiner@bifie.at
Silke Schwaiger, BIFIE Salzburg, s.schwaiger@bifie.at
Martin Seitz, BG/BRG Mödling Keimgasse, m.seitz@ph-noe.ac.at
Günther Sigott, Alpen-Adria-Universität Klagenfurt, guenther.sigott@aau.at
Klaus Siller, PH Salzburg Stefan Zweig, klaus.siller@phsalzburg.at
Michael Sörös, Stadtschulrat für Wien, michael.soeroes@ssrwien.gv.at
Carol Spöttl, Universität Innsbruck, c_spoettl@hotmail.com
Jürgen Struger, Alpen-Adria-Universität Klagenfurt, juergen.struger@aau.at
Michael Themessl-Huber, BMBWF, Wien, michael.themessl-huber@bmbwf.gv.at
Theresa Weiler, BMBWF, Wien, theresa.weiler@bmbwf.gv.at
Irmtraud Weyrich-Zak, BMBWF, Wien, irmtraud.weyrich-zak@bmbwf.gv.at
Matthias Zehentner, Universität Innsbruck, matthias.zehentner@uibk.ac.at