

University of Szeged, Faculty of Arts
Graduate School of Educational Sciences

TIBOR VÍGH

**A SYSTEM LEVEL ANALYSIS OF EXAM FUNCTION
AND WASHBACK EFFECT: STUDENT PERFORMANCE
AND TEACHER VIEWS ON THE HUNGARIAN MATURA EXAM
IN FOREIGN LANGUAGES**

Summary of the PhD Dissertation



Supervisor: Dr. Tibor Vidákovich

Doctoral programme
in Educational Evaluation

Szeged, 2010

The role of educational evaluation in educational systems and its feedback loops are finely defined by system-level models of teaching and learning informed by cybernetics (Nagy, 1979; Báthory, 1985). Innovations and reforms for educational change can always be traced back to the recognition of the importance of feedback. Exams functioning adequately in public education play an important role in these innovation processes and may also act as control mechanisms for the educational system, possibly having a strong impact on the teaching and learning process. Accordingly, the change of assessment and evaluation is an old reform strategy to initiate changes in education (Noble and Smith, 1994). Besides these reforms that initiate changes, feedback from the analysis of performance on exams and the exploration of the experience of participants of the teaching-learning processes involved may also contribute to the improvement of exams.

The knowledge of students can be represented by numerical exam scores, which may influence the academic progress of students, including their transition between levels of education. Exam results may serve as basis for decisions aimed at optimising instruction processes and provide a baseline for evaluating the realisation of the objectives set. In addition, the assessment and interpretation of the experience of educators may provide valuable feedback for exam development. Such information shows exam developers what impact the given exam has on the teaching and learning process and whether this impact corresponds to the original intentions.

Exams affect classroom processes and may induce a modernisation of content and methodology. The reason behind this is that teaching for the exam becomes the focus of teachers' thinking and daily practice, thus determining the how students study and influencing their efficacy. This feedback mechanism, known for decades and related issues are often discussed when new assessment and evaluation systems, such as school leaving and entrance exams are introduced or modified. Therefore, it is necessary to learn about the processes induced by exams, the functioning of exams and the factors influencing exams in detail, and, on this basis, to construct a model of this phenomenon.

THE PRECURSORS AND THE OBJECTIVES OF THE PRESENT DISSERTATION

The investigation of the above issues is relevant for the two-tier Matura introduced in Hungary in 2005, too (for a description, see Horváth and Lukács, 2006). This exam is seen as having great impact on public education and as initiating a modernisation of the contents and methodology of the teaching and learning process (Kósa and Simon, 2006). However, little is known about the actual impact of the Matura, and the relationship of the two tiers. In addition, the content standards need to be improved, and the two exam tiers need to be linked to performance levels identified in international assessment programs (Csapó, 2009). Monitoring the results, controlling the operation of exam functions, and exam development based on information from these sources are not strong characteristics of the current Matura. An important question to investigate is how this exam should be improved in order for its measurable impact to match the original intentions of its creators.

The dissertation focuses on the Matura in foreign languages, a compulsory subject of this exam, in order to demonstrate the role of assessment and evaluation in exam development and also the impact of the exam on public education as reflected in teachers' beliefs. Four considerations justify the choice of foreign language Matura exams as the subject of the present studies. First, need for the knowledge of foreign languages grew significantly in the 1990s, and the prestige of competent language use has become even more marked since Hungary joined the European Union. Second, as a result of the process in which the foreign

language Matura was shaped (see *Alderson, Nagy and Öveges, 2000; Einhorn, 2004*), the exam corresponds to the model and levels of communicative language competencies of the *Common European Framework of Reference (CEFR, 2001)*. Third, this exam has a place in the system of language exams, because the advanced level exam result is accepted as an equivalent to a language certificate. Fourth, the theory and practice of foreign language assessment and evaluation has been transformed significantly both internationally and in Hungary. Examples of changes include the emergence of competency assessment systems and the principles of communicative language testing, which characterise large sample foreign language assessment surveys and the foreign language Matura since the 2000s.

The foreign language surveys targeting Hungarian public education and the Matura correspond to international trends in research on language testing from three aspects. First, applying the principles of communicative testing connects them to current issues in language assessment and evaluation. Second, the validation of the levels of language proficiency tests against the CEFR levels is a field receiving great attention in European language assessment projects, similarly to building foreign language test banks made possible by this validation. Third, several studies analyse the effects of communicative testing on language teaching and learning (known as the washback effect in applied linguistics) and propose models of the washback processes. The review of the literature identified several issues that should be studied empirically in the Hungarian context, too, which, however, have received no or negligible research attention here in the past decade.

The objective of the dissertation is to describe the international tendencies of language assessment and evaluation and to identify their Hungarian characteristics, enabling a synthesis of the most important international and Hungarian conclusions in a theoretical model. The purpose of modelling language assessment and evaluation mechanisms is to provide a complex and multiperspective analysis of the Matura and to validate the model empirically. The findings and conclusions outline tasks regarding the improvement of the Matura, further research and in-service training for the exam.

RESEARCH QUESTIONS AND METHODS

I. The analysis and synthesis of the literature aims to identify the international trends of language assessment and evaluation. Research questions in this area included:

- (1) how communicative testing integrates into test theory as a field and how the qualitative criteria of competence assessment surface in the assessment and evaluation of communicative competence;
- (2) which testing methods are applied in the practice of communicative testing.

As regards the effects of communicative testing on language instruction, the literature review focused on

- (3) how washback can be interpreted and characterized;
- (4) what factors influence the washback process.

The analysis of the two-tier foreign language Matura exam requirements targeted

- (5) the examination of the fit between the exam and the specified model of language competence on the one hand, and the qualitative criteria of communicative testing on the other hand;
- (6) the formulation of hypotheses regarding the washback effect of the exam, based on the findings of international studies.

The analysis of the international trends and the synthesis of Hungarian characteristics enabled

- (7) the construction of a model of language assessment and evaluation mechanisms.

The development of different washback models was traced, and the factors influencing washback were identified in the literature. This review provided the basis for the construction of a synthesised model, which, in turn, was exemplified with the characteristics of the Matura. The elements of the model to be analyzed in the dissertation were selected and justified with regard to the validation of this synthesised model. The empirical studies presented in the dissertation focused on the current functioning of the Matura. Its internal and external validity and the quality of assessment and evaluation processes of writing skills were investigated through the analysis of the exam results. Feedback from teachers provided the data base for examining the washback effect of the exam on the teaching and learning process.

II. Performances at recent foreign language Matura exams were used to generate information to promote the improvement of the exam and to offer feedback on needs and objectives by analysing

- (8) how performances at the foreign language Matura differ in individual subtests;
- (9) what differences can be found between Matura examinees regarding the levels of their language abilities by school type, instructional programme and gender, when comparing of English and German scores.

These research questions were answered using the free access database of the Oktatási Hivatal (OH, Hungarian Educational Authority). In the first phase of data analysis, performances were described, then the differences in the developmental levels of language skills were mapped and their impact on performance was analysed for both languages.

III. The prime reason for the examination of the fit between sub-tests of receptive skills and students' language proficiency on the one hand, and the CEFR levels (2001) on the other hand, came from the intention to provide feedback for the developmental process of the exam, and to raise the quality of exam functioning. Research questions in this section include:

- (10) to what extent is the ability range of examinees covered by the empirical difficulty of the Matura exam tasks;
- (11) how can a language test bank based on the CEFR levels be created.

In the first phase, the 2006 ordinary level German as a foreign language Matura items in reading comprehension, listening comprehension and use of German sub-tests were analysed using the Rasch model (*Rasch*, 1960). The three subtests were first calibrated to one common scale, then the difficulty indices of the tasks were represented in three dimensions in a person-item map. In the second phase, the person-item map of the ordinary and advanced level sub-tests of 2005 and 2006 were analysed individually. Furthermore, the results and findings of several European projects were also studied (e.g. *CITO*, 2008).

IV. The investigation of the assessment and evaluation processes of the writing skills sub-test was aimed at defining quality criteria and parameters that enable the comparison of results from different years and make it possible to trace changes over time. Therefore the analysis focused on

- (12) defining parameters to characterize the assessment criteria and rating scores;
- (13) showing how the analytical scales of assessment criteria function.

This study used the database of the advanced level German language Matura of 2006. First, the data were calibrated to a common scale with the partial credit model (*Masters*, 1982) to examine the difficulty of evaluation criteria, which was linked to examinees' proficiency in writing. Next, the probability curves of the scales were studied to examine raters' judgements. The applicability of the parameters identified was also analyzed with data from the ordinary level German Matura exam.

V. The objective of studying feedback from teachers of English and German was to understand the perspective of teachers involved with the Matura, i.e. how they see the functioning of the exam from as well as the washback of the Matura on the teaching and learning process. Research questions formulated in this section included:

- (14) What beliefs and attitudes do language teachers have concerning the exam and what exam-related experiences affect them?
- (15) What factors determine language teachers' instructional practices and how do exam requirements and tasks shape these factors?
- (16) Do teachers' beliefs and attitudes towards the exam determine the content, methods, and assessment and evaluation practices of language instruction?
- (17) To what extent does the foreign language Matura promote the dissemination of the assessment and evaluation methods of communicative testing in the teaching and learning process?

A questionnaire was administered to 149 language teachers. The sample is representative for the foreign languages taught and for school type. The literature was consulted when developing the instrument as regards the structure of the questionnaire as well as the factors and questions to be included (for example *Spratt, 2005; Cheng, 1998; Chen, 2002*). The reliability of the questionnaire is adequate (Cronbach $\alpha=.89$). Its validity was confirmed by factor analysis. The analysis first focused on teachers' beliefs and attitudes concerning the exam, then on the washback effect of the foreign language Matura was mapped in teachers' beliefs. Next the strength of the impact of beliefs and attitudes on the process of language teaching and its assessment and evaluation processes was determined.

RESULTS AND FINDINGS

Communicative testing in the system of foreign language assessment and evaluation

(1) Communicative language testing is an independent and specific field of educational evaluation that relies on classical test theory to determine the quality criteria, the functions and the types of communicative tests (*Bachman, 1990*). However, the characteristics of communicative language competencies lead to the appearance of specialised meanings in the concepts of assessment methodology and assessment and evaluation processes of general test theory. Communicative language testing also enriches the criteria of classical test theory, highlighting the importance, and defining new types and aspects of validity (*Messick, 1996*). The latter include the authenticity, interactivity and effectiveness of the test development process (*Bachman and Palmer, 1996*); and also the fairness, comparability, reproducibility of decisions (*Kunnan, 2004*) and washback (*Alderson and Wall, 1993*) in administering the test and its consequences. These criteria are in line with the quality criteria of competence assessment (*Baartman et al, 2006*).

(2) Of special importance among test types are competence and performance tests (*McNamara, 2000*), as well as the closely related types of direct and indirect tests, each pair representing the two end points of a scale. Test developers must find a balance between these tests types (*Alderson, Clapham and Wall, 1995*). These methods of testing are used in language exams, a process also promoted by *CEFR (2001)*. Thus it is an important issue how exam developers can fit tasks and evaluation processes to the CEFR levels. In this process, models of item response theory can be applied in order to enhance efficacy.

The washback effect of communicative testing on teaching and learning

(3) There is a lack of unified terminology for the concept encompassing the impact of language exams, with researchers defining the phenomenon in different ways and attributing different conceptual contents to the construct (*Bailey, 1999*). Having synthesised these concepts, washback appears to be a complex phenomenon exerting both direct and indirect effects on people involved in the exam, on the teaching and learning process, and on the

content and quality of learning (Alderson and Wall, 1993). In a broader sense, it also bears on public education and society through the decisions made based on the results of the exam (Hamp-Lyons, 1997).

(4) Models emerging from theoretical research so far have not given precise descriptions of the characteristics of process components and their interaction with the exam results (Wall, 2005). In these models, emphases fall on needs and objectives (Saif, 2006), the importance of the prestige exams in inducing the effects (Watanabe, 2004), and the crucial role of teachers' beliefs, assumptions, knowledge and attitudes (Burrows, 2004). The findings of empirical studies (e.g. Cheng, Watanabe and Curtis, 2004) can only be interpreted in their given context and are difficult to generalise. Still, it can be established that, in general, the exam has greater a washback effect on the content of teaching than on the methods of instruction (Alderson, 2004), though there is little research on the effects on classroom evaluation.

The two-tier Matura in the system of communicative testing

(5) The two-tier foreign language Matura is intended to be authentic, interactive, effective and fair. Evidence supporting these claims of the exam was obtained from the analysis of its requirements, and the conformity of these to the CEFR model of language competencies explicitly targeted in the Matura, and also from the comparison with the criteria, functions and types of communicative testing. However, mainly because exam functions are confused, and because the monitoring of exam quality is lacking, it is dubious whether comparability and the reproducibility of decisions are achieved. The Matura relies primarily on the use of weak language performance tests.

(6) By assessing communicative language use, the two tier foreign language Matura takes language learners' needs into account. It is highly prestigious, because it fulfils the functions of a school leaving exam, an entrance exam to higher education and a language certificate. All factors are present that research has identified for inducing the washback mechanisms of the exam, therefore the Matura may forcefully shape teachers' beliefs and attitudes to the exam, which, in turn, may influence the process of language teaching and affect teachers' beliefs related to assessment and evaluation processes.

The synthesised model of washback functions

(7) The model created on the basis of the literature review serves as the basis of the empirical studies presented in the dissertation. It includes the representation of the washback mechanism of language exam on the teaching and learning process. According to this model, the exam first affects educators' beliefs and attitudes, which, in turn, exert influence on the instructional process as well as the content and quality of learning. The model includes two further feedback cycles of different types. One stems from exam results and feeds into the exam development process, regarding the changes in objectives and needs. The other cycle draws on surveys on the beliefs and attitudes of those involved in the teaching and learning process, feeds into exam development processes and facilitates the adequate interpretation of results. The empirical research presented in the dissertation was designed to correspond to this model in order to examine the foreign language Matura as reflected in exam results and the opinions of language teachers.

The current functioning of the exam

(8) The exam results show that, in both languages, performances at the ordinary level Matura are lower on the writing skills and use of German/English sub-tests, and higher on the reading and listening comprehension and oral subtests. For both languages, average

performances vary on the reading comprehension and listening comprehension sub-tests. This signals that the performance on different test versions varies, therefore the reproducibility of decisions is low. At the advanced level, average performances are high in both languages, the majority of examinees performing in the 80 to 100% range except for the use of German/English subtest. Therefore this exam does not fulfil adequately the requirements of an entrance exam to higher education for foreign language majors.

(9) Differences by school type are significant at the ordinary level and show continual annual increase. Similar tendencies were found for both languages, i.e. those taking the Matura having completed their secondary education in the academic stream perform significantly better than those coming from technical schools at both exam levels. Also, full time students in secondary education perform significantly better than those who take part in adult part-time secondary education. Regarding productive skills, girls outperform boys, however, the differences by gender do not have a significant impact on overall exam performance.

The internal and external validity of the exam

(10) The task and item level analysis of sub-tests assessing receptive skills in the 2006 ordinary level German Matura showed that the difficulty order determined by task developers usually matched the increase in empirical difficulty. The item difficulty parameters of exam tasks do not cover fully the scale of skill levels exhibited by examinees, however, this finding cannot be generalised to all sub-tests. The items in the use of German tasks cover mostly accurately examinees' ability range. As regards reading and listening comprehension tasks, two groups of items could clearly be distinguished, one assessing examinees below the average ability level, and the other assessing those above it. In addition, a high ability level was identified which these tests do not assess.

(11) Findings from the analysis of the individual person-item maps of each skill yielded arguments for the necessity of examining the equivalence of tests versions, for the analysis of the relationship of the two Matura levels to each other and to the CEFR levels in rounds of new data collection with anchor and reference items, and for the consequent creation of an IRT based item bank. Findings from European projects may be useful in selecting reference tasks and in characterising texts and items for the Matura tasks. The proposed item bank could function as a reference point for the development of exam tasks and could contribute to the empirical verification of the correspondence between the Matura and CEFR language proficiency levels.

The quality of assessment and evaluation procedures of writing skills

(12) The results of the writing skills sub-test of the Matura exam in 2006 in German as a foreign language provide the basis for the characterisation of assessment criteria regarding their average difficulty level and their fit mean square index. Misfitting assessment criteria may indicate that they assess other dimensions of writing skills (e.g. the first task of the ordinary level exam) or they function in different ways (e.g. the communicative components at both levels). The functioning of rating scores is accurate if their thresholds increase proportionally to the rising of writing skill levels. This was found to be the case at the exam, but disproportionality was detected for the assessment criteria of text production and of vocabulary and expression. The high value of the fit mean square index of the rating scores may indicate that raters give the same scores to examinees of different ability levels. At the advanced level, this was found primarily in the higher ability ranges for grammar and spelling. The person-item map revealed the existence of a higher proficiency level of writing skills which is not covered by the thresholds of rating scores.

(13) The analysis showed that the functioning of the rating scales differs by criterion. Rating scale use was characterised by probability curves which showed the redundant functioning of the scales (e.g. in text production) and indicated that raters do not use certain rating scores (e.g. regarding content).

The washback effect of the exam on the language teaching process as reflected in teachers' feedback

(14) The results show that teachers consider the present foreign language Matura exams acceptable as well as more reliable and more valid as the previous paradigm. The teachers' examination experiences prove to be crucial in shaping their beliefs about the exam.

(15) Teachers report that they see the practice of foreign language instruction as influenced significantly by previous experience in foreign language teaching and the Matura exam requirements. Familiarising themselves with its basic principles as discussed in Matura documents, teachers attribute greater importance to the criteria of communicative testing when preparing their own assessment tasks.

(16) In accordance with international research findings on washback, Hungarian teachers were found to attribute a larger effect to the exam on the content of foreign language teaching than on its methods.

(17) It was expected that the principles of evaluation of the Matura affect teachers' beliefs about assessment and evaluation procedures. This hypothesis was verified in the Hungarian contexts as the findings showed that teachers' beliefs about, and attitudes to, the Matura influenced several aspects of classroom assessment and evaluation. For example, they considered the criteria of communicative testing more important and they were more inclined to use exam tasks properly, corresponding to their original functions.

CONCLUSIONS AND SUGGESTIONS

Finally, the conclusions that can be drawn from the results of our survey are summarised and suggestions for further analyses are presented.

The analysis of the current functioning of the exam at the ordinary level revealed significant differences in the language proficiency of examinees by school and programme type, and showed that these differences have increased over the years. At the advanced level, main problem areas include requirement levels, the quality of exam tasks and the accomplishment of the function of the Matura as entrance exam to higher education. These areas indicate further needs regarding exam development and also new objectives for development. It seems necessary to identify and diagnose the causes of the differences in achievement and to monitor and revise the quality of assessment and evaluation processes. Since the data bases of the Hungarian Educational Authority are not suitable for this purpose, test and item level analyses are needed.

In examining the internal and external validity of the exam, the model of item response theory was used, therefore the findings can be generalised from a methodological point of view. The empirical validation of requirement levels is an important research task in other exam subjects as well. In order to raise the quality of assessment and the reproducibility of decisions, it would be important e.g. to link the difficulty of tests in mathematics, science and reading comprehension to the proficiency levels of international assessments (e.g. the PISA surveys). It would also be important to identify the differences between the two Matura levels concerning content and quantity and to create a test bank.

The analysis of the writing skills sub-test was presented as an example for the empirical validation of assessment criteria and scales. The parameters identified in this sub-test make it

possible to compare exam results and to trace their changes. These analyses also provide an adequate basis for the assessment of oral skills in foreign languages, the sub-test of which exhibits low reliability at the Matura. Moreover, the methods used and the quality criteria presented can also be appropriated, applied and analysed in exam subjects and surveys where assessment is based on analytic scales.

The teacher survey focuses the attention of exam developers on the importance of exam-related in-service training programmes for familiarising teachers with the philosophy of the Matura. At the same time, this study has several constraints. No classroom-level analysis could be attempted here regarding the effects of Matura in foreign languages, and no information can be drawn regarding the relationship between students' learning processes and the Matura. Thus it is not fully understood how the exam exerts its influence on the teaching and learning process. Research on instructional methodology is needed by subject fields to provide a detailed analysis of teachers' instructional practices and students' learning processes.

The studies reported in the dissertation clearly outline the necessity of exam development and further research. Its relevance confirmed by empirical evidence to a great degree in the research presented, the theoretical model proposed here appears to be useful and appropriate in planning this development and research, identifying the domains to be targeted, and to characterise the relationships between these domains. However, the validity of the findings is restricted, as they primarily concern the Matura exam in foreign languages. Studying all the elements of the model is beyond the scope of this dissertation and the elements not discussed here would require a long-term research design. Therefore further studies should be done that would enable the Matura to develop into an annual, regular system-level assessment which could yield feedback for those involved in public education on the one hand, and make it possible to monitor the knowledge of examinees over the years on the other hand.

REFERENCES

- Alderson, J. C. (2004). Foreword. In: Cheng, L., Watanabe, Y. & Curtis, A. (Eds.). *Washback in language testing: Research contexts and methods* (pp. ix–xii). Mahwah, NJ.: Lawrence Erlbaum Associates.
- Alderson, J. C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Alderson, J. C. Nagy, E. & Öveges, E. (2000, Eds.). *English language education in Hungary. Part II. Examining Hungarian learners' achievements in English*. Budapest: The British Council Hungary.
- Alderson, J. C. & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2) 115–129.
- Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A. & van der Vleuten, C.P.M. (2006). The wheel of competency assessment. Presenting quality criteria for Competency Assessment Programs. *Studies in Educational Evaluation*, 32(2) 153–177.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bailey, K. M. (1999). *Washback in language testing. TOEFL Monograph Series, Ms. 15*. Princeton, NJ.: Educational Testing Service.
- Báthory Z. (1985). *Tanítás és tanulás*. Budapest: Tankönyvkiadó.
- Burrows, C. (2004). Washback in classroom-based assessment: A study of the washback effect in the Australian adult migrant English program. In: Cheng, L., Watanabe, Y. & Curtis, A. (Eds.). *Washback in language testing: Research contexts and methods* (pp. 113–128.). Mahwah, NJ.: Lawrence Erlbaum Associates.
- CEFR (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: University of Cambridge.
- Chen, L. (2002). *Taiwanese junior high school English teachers' perceptions of the washback effect of the basic competence test in English. Doctoral Dissertation*. Ohio: The Ohio State University.
- Cheng, L. (1998). *The washback effect of public examination change on classroom teaching: an impact study of the 1996 Hong Kong Certificate of Education in English on the classroom teaching of English in Hong Kong secondary schools. Doctoral Dissertation*. Hong Kong: University of Hong Kong.

- Cheng, L., Watanabe, Y. & Curtis, A. (2004, Eds.). *Washback in language testing: Research contexts and methods*. Mahwah, NJ.: Lawrence Erlbaum Associates.
- CITO (2008). Building a European bank of anchor items for foreign language skills. EBAFLS. General report. Arnheim: Cito B.V.
- Csapó, B. (2009). The second stage of public education and the Matura. In: Fazekas, K., Köllő, J. & Varga, J. (Eds.). *Green book for the renewal of public education in Hungary* (pp. 81–104). Budapest: Ecostat.
- Einhorn, Á. (2004, Eds.). *Az érettségiről tanároknak 2005. Német nyelv*. Budapest: Országos Közoktatási Intézet. (http://www.okm.gov.hu/letolt/kozokt/erettsegi2005/tanaroknak/nemet/nemet_nyelv.htm, 17. 03. 2006.)
- Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing*, 14(3) 295–303.
- Horváth, Zs. & Lukács, J. (2006, Eds.). *Új érettségi Magyarországon. Honnan, hová, hogyan? Egy folyamat állomásai*. Budapest: Országos Közoktatási Intézet.
- Kósa, B. & Simon, M. (2006, Eds.). *Új vizsga – új tudás? Az új érettségi hatása az iskolakezdéstől a záróvizsgáig*. Budapest: Országos Közoktatási Intézet.
- Kunnan, A. J. (2004). Test fairness. In: Milanovic, M. & Weir, C. (Eds.). *European language testing in a global context* (pp. 27–48). New York: Cambridge University Press.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2) 149–174.
- McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3) 241–256.
- Nagy, J. (1979). *Köznevelés és rendszerszemlélet*. Veszprém: OOK.
- Noble, A. J. & Smith, M. L. (1994). *Measurement-driven reform: Research on policy, practice, repercussion*. CSE Technical Report 381. Arizona: Tempe, Arizona State University.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Language Testing*, 23(1) 1–34.
- Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. *Language Teaching Research*, 9(1) 5–29.
- Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching*. *Studies in Language Testing* 22. Cambridge: Cambridge University Press.
- Watanabe, Y. (2004). Teacher factors mediating washback. In: Cheng, L., Watanabe, Y. & Curtis, A. (Eds.). *Washback in language testing: Research contexts and methods* (pp. 129–146). Mahwah, NJ.: Lawrence Erlbaum Associates.

PUBLICATIONS RELATED TO THE DISSERTATION

- Nikolov M., Tóth E. & Vígh T. (2009). The impact of background factors on Hungarian students' performances on English and German reading comprehension tests. In: Riljaarsdam, G. (Eds.). *Forstoring Communities of Learners. 13th Biennial Conference of the European Association for Research on Learning and Instruction* (Abstracts, p. 180). Amsterdam: Graduate School of Teaching and Learning, UvA.
- Vidákovich, T. & Vígh, T. (2009). Az idegen nyelvi értékelés és oktatás kapcsolata. In: Kozma, T. & Perjés, I. (Eds.). *Új kutatások a neveléstudományban 2008. Hatékony tudomány, pedagógiai kultúra, sikeres iskola* (pp. 159–168). Budapest: Magyar Tudományos Akadémia Pedagógiai Bizottsága.
- Vígh, T. (2005). A kommunikatív tesztelés elméleti alapjai. *Magyar Pedagógia*, 105(4) 381–407.
- Vígh, T. (2006). Der Prüfungsteil Schreiben im neuen Abitur für DaF. In: Orosz, M. & Albrecht, T. (Eds.). *Jahrbuch der ungarischen Germanistik 2005* (pp 335–351). Budapest & Bonn: Gesellschaft Ungarischer Germanisten & Deutscher Akademischer Austauschdienst.
- Vígh, T. (2007a). A kommunikatív tesztelés az idegen nyelvi mérés és értékelés rendszerében. In: Korom, E. (Eds.). *PÉK 2007 – V. Pedagógiai Értékelési Konferencia: Program – Tartalmi összefoglalók* (Abstracts, p. 78). Szeged: SZTE, Neveléstudományi Doktori Iskola.
- Vígh, T. (2007b). A vizsgák tanítási-tanulási folyamatra gyakorolt hatásának elméleti és empirikus kutatása. *Magyar Pedagógia*, 107(2) 141–161.
- Vígh, T. (2008a). A nyelvi szintek kialakításának és nyelvi feladatbank kiépítésének vizsgálata az idegen nyelvi érettségi vizsgán. In: Perjés, I. & Ollé, J. (Eds.). *VIII. Országos Neveléstudományi Konferencia: Hatékony Tudomány, Pedagógiai Kultúra, Sikeres Iskola. Program és Összefoglalók* (Abstracts, p. 149). Budapest: Magyar Tudományos Akadémia, Pedagógiai Bizottság.
- Vígh, T. (2008b). A vizsgák tanítási-tanulási folyamatra gyakorolt hatásának modellezése. In: Csíkos, Cs. (Eds.). *PÉK 2008 – VI. Pedagógiai Értékelési Konferencia: Program – Tartalmi összefoglalók* (Abstracts, p. 97). Szeged: SZTE, Neveléstudományi Doktori Iskola.

- Vígh, T. (2008c). Egy IRT-alapú nyelvi feladatbank létrehozásának módszertani kérdései. A német nyelvi érettségi vizsgafeladatok eredményei. *Magyar Pedagógia*, 108(1) 29–51.
- Vígh, T. (2008d). Theoretische und empirische Erforschung von Rückwirkungsmechanismen des Abiturs für DaF. In: Tichy, E. & Masát, A. (Eds.). *Jahrbuch der ungarischen Germanistik 2007* (pp. 302–317). Budapest & Bonn: Gesellschaft Ungarischer Germanisten & Deutscher Akademischer Austauschdienst.
- Vígh, T. (2009a). Az íráskészség mérési-értékelési eljárásainak minőségi mutatói az emelt szintű német érettségiben. In: Bárdos, J. & Sebestyén, J. (szerk.). *IX. Országos Neveléstudományi Konferencia: Neveléstudomány - Integritás és integrálhatóság. Program és tartalmi összefoglalók*. (Abstracts, pp. 210–211). Veszprém: Magyar Tudományos Akadémia, Pedagógiai Bizottság.
- Vígh, T. (2009b). Der Prüfungsteil Schreiben im neuen ungarischen Abitur für Deutsch als Fremdsprache. In: Hunstiger, A. & Koreik, U. (Eds.). *Chance Deutsch: Schule – Studium – Arbeitswelt. 34. Jahrestagung des Fachverbandes Deutsch als Fremdsprache 2006 an der Leibniz Universität Hannover. Materialien Deutsch als Fremdsprache Band 78* (pp. 149–165). Göttingen: Universitätsverlag Göttingen.
- Vígh, T. & Vidákovich, T. (2008). Az idegen nyelvi tesztelés és oktatás kapcsolata: a nyelvtudást mérő vizsgák visszahatásának elméleti és empirikus kutatása. In: Perjés, I. & Ollé, J. (Eds.). *VIII. Országos Neveléstudományi Konferencia: Hatékony Tudomány, Pedagógiai Kultúra, Sikeres Iskola. Program és Összefoglalók* (Abstracts, p. 148). Budapest: Magyar Tudományos Akadémia, Pedagógiai Bizottság.