CSP **Center scientific da cumpetenza per la plurilinguitad** Cogniziun Società Formation Bildung Migration Furmaziun Gesellschaft
CSP **Centro scientifico di competenza per il plurilinguismo** Scuola Arbeit Politique Communitad School Travail Ecole Community
CSP **Centre scientifique de compétence sur le plurilinguisme** Migrazione Furmaziun Societad Cognition Society Scola Migration
KFM **Wissenschaftliches Kompetenzzentrum für Mehrsprachigkeit** Societé Cognizione Migraziun Schule Communauté Kognition
RCM **Research Centre on Multilingualism** Formazione Lavoro Politics Comunità Work Politik Lavur Politica Formation Gemeinschaft

Language requirements and language testing for immigration and integration purposes

A synthesis of academic literature

Evelyne Pochon-Berger
Peter Lenz

2014

Report of the Research Centre on Multilingualism

INSTITUT FÜR MEHRSPRACHIGKEIT
INSTITUT DE PLURILINGUISME
ISTITUTO DI PLURILINGUISMO
INSTITUT DA PLURILINGUITAD
INSTITUTE OF MULTILINGUALISM

HEP | PH FR

UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Published by
Institute of Multilingualism
www.institute-multilingualism.ch

Authors
Evelyne Pochon-Berger
Peter Lenz

This project was realised as part of the Research Centre on Multilingualism's 2012–2014 programme and received funding form the Swiss Confederation. The opinions expressed in this work are solely those of the authors.

Fribourg, March 2014

Layout
Billy Ben, Graphic Design Studio

Language requirements and language testing for immigration and integration purposes

A synthesis of academic literature

Evelyne Pochon-Berger
Peter Lenz

2014

Report of the Research Centre on Multilingualism

Abstract

The present document is the result of a literature-based study of recent developments in the field of language requirements and language testing for immigration and integration purposes. The study was carried out as part of the regular research programme 2011-2014 at the Swiss national Research Centre on Multilingualism (RCM)[1].

The introduction of language requirements and the use of formal language tests as prerequisites for obtaining entry visas, residence permits and/or citizenship are fairly recent but rapidly evolving practices in many Western countries, particularly in Europe. These requirements and assessments have been made the object of academic debate and research, and the subject of several scholarly publications. This report seeks to present a structured overview of several aspects of these requirements and assessments; the basis for our study is provided by a corpus of academic publications.

The summary of literature put forward in this paper is divided into two main chapters. The first provides an overview of recent developments, mainly over the past decade, regarding language requirements for immigration, residency and naturalization; it also describes the introduction of new, formal assessments at different stages of the immigration process. Because the information presented is often based on accounts by academics critical of the new tendencies, this chapter also introduces several contested issues, primarily from a critical perspective. The second main chapter deals systematically with topics from the ongoing debate on language testing for immigration and integration purposes under consideration of major categories from assessment validation research, with Bachman & Palmer's (2010) Assessment Use Argument (AUA) framework providing the conceptual background for our account.

Our study concludes with a series of recommendations for future assessment-related research, publications, and assessment validation practices.

Overall, our findings confirm the subjectively perceived increase in language requirements and formal language testing (sometimes complemented by cultural knowledge testing) for immigration and integration purposes in several 'Western' countries, particularly in Europe, over the past decade. Although a common general trend is discernible, we are also able to observe considerable diversity in the concrete definition of the requirements and the corresponding assessments, for instance, regarding the levels of language proficiency required.

The analyses of the literature on the basis of Bachman and Palmer's validation framework reveal that the published literature treats the various topics involved rather unevenly. In particular, only very few papers are actually aimed to systematically demonstrate (or challenge) the validity of a specific test and its use. Most papers deal with either the *meaningfulness* of test content and levels with respect to the knowledge and skills selected for assessment (typically 'sufficient knowledge to integrate in society') and the *beneficence* a test and the preparatory phase leading up to it hold for various stakeholders, or they raise questions about the testing of immigrants *per se* which are so fundamental in nature that they are only marginally accommodated by existing validation frameworks.

In our conclusion, we make recommendations for future work concerning the following points: the inclusion of academic disciplines specialising in social, ethical and legal topics related to language testing for immigration and integration purposes; a commitment to accountability and systematic test validation within an up-to-date validation framework; the employment of needs analyses as a basis for test design and use; the need for more impact studies on test regimes; the desirability of more transparency concerning test validation and its outcomes.

1 | The authors would like to thank Séverine Beaud Duarte Rodrigues for the great contribution she made to the collection and screening of academic publications in the early phase of the project. The valuable comments of our internal reviewer, Thomas Studer, made at different stages of the study are also greatly appreciated.

Contents

1 Introduction

1.1 The Swiss context

In many Western countries, particularly in Europe, the authorities increasingly link their decisions on the residency status of non-citizens to language tests. Switzerland is not at the forefront of this development, but recent efforts by legislations and administrations at different levels of the political system indicate a tendency in this direction. At the federal level, language issues are treated as an aspect of integration. Under the official motto guiding integration policy, 'fördern und fordern' ('encourage/support and demand'), the authorities launched a master plan for integration in 2010 that stipulates increased language support for foreign speakers of languages that are not one of the national languages of Switzerland (Gerber & Schleiss, 2013). In a first phase, a number of instruments were developed to encourage and support language teaching and learning, such as a framework curriculum for 'low-threshold' language courses (Lenz et al., 2009), needs-oriented specifications of objectives based on scenarios of language use (Müller & Wertenschlag, 2013), or materials for needs-oriented student placement, course planning, teaching and (self-)assessment. In addition, a workshop programme was launched to facilitate the dissemination of the new instruments.

Since 2008, the Law on Foreigners has allowed the authorities to define language requirements (usually attendance of a language course) as part of a so-called integration agreement with persons applying for residency (Skenderovic, 2013). Specific language assessments (including tests[2]) for residency[3] do not yet exist. At the time this report was written (2013) and the revised Law on Foreigners was in the legislative process, the 'demand' part of the official motto was receiving increasing attention. It is likely that, from 2015 onwards, certain groups of immigrants will have to supply evidence of their knowledge of a national language, most probably by passing either an officially recognised test or officially controlled course-related assessments.

1.2 Purpose and structure of this report

The main purpose of this study is to provide a structured overview of several aspects related to language assessment for immigration and integration purposes as seen through the lens of academic publications. These aspects are:

- the recent developments regarding language requirements and formal language testing for immigration and integration purposes;
- the issues at stake and the arguments in the primarily critical discussions revolving around these developments;
- the validity of language testing and tests for immigration and integration purposes.

By providing an overview, we hope to identify not only current issues and major trends, but also to point out shortcomings in current academic research and debate. This information then serves as a basis for making recommendations in view of further conceptual work, research and development in this sensitive area.

The present study has a straightforward structure: chapter 1 provides the background for the study, that is, the local context, purpose and structure of the study as well as published sources available. Chapter 2 informs about the developments over the past decade as pertains to language requirements for immigrants; it also describes the introduction of new formal assessments at different stages of the immigration process (immigration, residency, and naturalisation). Because this information is based on accounts by academics predominantly opposed

2 | 'Assessment' is a more open term than 'test'. It covers tests as well as other activities that help establish the degree of knowledge or skill of the individual undergoing assessment. Whenever an assessment happens to have the form of a test, the two terms are used interchangeably in this document.

3 | Citizenship is a separate issue; the communes (still) have the right to grant it and to define the language requirements and the assessment procedures. There is considerable variation in the practices adopted.

to the introduction of more stringent conditions for immigration, chapter 2 also provides a structured overview of the various points of criticism. Chapter 3 begins by outlining a widely accepted framework for test validation. The main categories from this framework are then used to present the current discussion on language testing for immigration and integration purposes. Finally, chapter 4 summarises the findings of the previous chapters, draws conclusions and makes recommendations for future work.

1.3 Sources

The use of defined language requirements and formal language tests as prerequisites for obtaining entry visas, residence permits and citizenship is a fairly recent development, but one that is rapidly evolving. Correspondingly, language testing for immigration and integration purposes is a new but growing field of scholarly research and debate that has attracted the attention of a number of academics. The studies that have emerged thus far are predominantly critical descriptions and analyses of language-related immigration policies and testing regimes introduced in various countries. Questions on test fairness and social justice of the new policies and systems have been widely addressed. Over the last decade, a number of edited volumes dealing with these topics appeared, such as Adami & Leclercq (2012), Hogan-Brun et al. (2009a), Mar-Molinero & Stevenson (2006), Oers et al. (2010), Slade & Möllering (2010). In addition a few special issues of journals were dedicated to these issues, e.g. *International Journal on Multicultural Studies*, vol. 10(1), 2008; *Language Assessment Quarterly*, vol. 6(1), 2009. A handbook article by Kunnan (2012) provides an up-to-date overview. Although most available publications are as yet mainly programmatic, they have nevertheless helped to open up the field for research and have raised the awareness applied linguists and language testers have of the recent developments and the issues involved. Published empirical studies, however, remain scarce, and very few studies focus on the development and validation of actual language tests. The articles by De Jong et al. (2009), Perlmann-Balme (2011) and Plassmann (2011) are exceptions. The paper by De Jong et al. (2009) has a distinct research orientation while the two other articles are accessible to a wider audience. All three papers share a confirmatory orientation – representing the test developers' views and interests – rather than taking a critical point of view.

As language testing for immigration and integration purposes is of no small concern for governments and inter-governmental institutions (such as the Council of Europe), and is furthermore a topic present in current public discourse and in politics, it is unsurprising that there are several officially mandated reports available on the topic (e. g. Schneider et al., 2006: recommendations concerning language testing for citizenship on behalf of the Swiss Federal Commission on Migration; Balch et al. (2008); and Extramania & Van Avermaet, 2010: results of two surveys conducted on behalf of the Council of Europe; Little, 2010: a report concerning the linguistic integration of adult immigrants, commissioned by the Council of Europe). Our study focuses on papers published in scholarly journals or books; only a small selection of commissioned reports, such as the aforementioned, are occasionally consulted.

The available literature is rather limited in its geographical scope. It covers mainly policies and practices in European countries (EU member states and non-EU states), and in only a few other countries, namely the USA, Canada and Australia. Papers were written on these countries in part because they have been facing considerable immigration flows. In addition, although some countries have a longstanding immigration tradition (e. g. USA, Canada, Australia and some Western European countries), for others this experience is a more recent development (e. g. Southern European countries such as Spain and other Western European countries) (Extra et al., 2009a). According to Van Avermaet (2009), drastic changes in immigration policies occurred in response to major increases in immigration flows. European countries are particularly concerned by the recent developments because immigration to Europe adds considerably to mobility within the European Union. The European context is therefore a propitious object for research. The great diversity of practices and policies – despite common tendencies – found across the European countries, has added to the research interest. There are studies available on the pioneering countries in matters of citizenship testing – the Netherlands, Germany, Denmark and the UK (see e. g. Blackledge, 2009a and b; Extra & Spotti, 2009a and b; Laversuch, 2008; Michalowski, 2010; Möllering, 2010; Slade, 2010a). Practices in traditional immigration countries overseas, such as the USA, Canada and Australia, are also well described (see e. g. Cox, 2010; Farrell, 2010; Holland, 2010; Hargreaves, 2010; McNamara, 2009a). Information on European countries such as Austria, Belgium, Luxembourg, Spain and the Baltic states is rather sparse, while practices in other countries remain largely unexplored (e. g. Greece, Italy, Poland, Portugal). Finally, we know close to nothing about the practices in non-Western countries (e. g. Asian or African countries; see Lee (2010) on the Korean context for an exception).

In the present paper, we direct our attention mainly to publications on European countries.

2

Language assessment for immigration and integration purposes in a changing context

2.1 Emerging language policies and assessment practices

In a number of countries, heated political debates have taken place on the topic of language testing and assessment for immigration and integration purposes; these debates have resulted in changes to national legislation. The following papers provide an overview: Blackledge (2009a & b) for the British context, Slade (2010a) for the Dutch context, Möllering (2010) for the German context, Farrell (2010) for Australia. At a supranational level, Balch et al. (2008) provides recommendations for the use of language testing for citizenship, and Extramania & Van Avermaet (2010) gives a summary of language policies on immigration in a number of European countries. A third report commissioned by the Council of Europe, forthcoming in 2014, will be based on a large-scale survey of member states' policies on the linguistic integration of immigrants that was carried out in 2013.

Some studies indicate a major shift in the national immigration policies of many countries (mainly in Europe but also Australia, Canada and the USA) over the past decade. This shift has materialised in an 'increasing number and extent of regulatory mechanisms' (Saville, 2009). In other words, there is a general tendency to implement stricter immigration and integration measures (Hogan-Brun et al., 2009b; Van Avermaet, 2009; Slade, 2010a), thereby revealing an 'assimilationist' approach to the treatment of immigrants (Möllering, 2010; Slade, 2010a). Some countries even set up sanction-oriented, compulsory integration programmes (Kostakopoulo, 2010). According to Böcker & Strik (2011), the tendency towards increasingly strict immigration policies in many European countries contrasts sharply with earlier policies adopted by these countries in the 1980s and 1990s. The former policies were geared towards promoting immigration and integration through easy terms for residence permits and naturalisation. The Netherlands, for example, used to be liberal in this respect, but it became the first European country to adopt strict immigration and integration conditions (Slade, 2010a). Since this reversal in legislation, the granting of permanent residency or citizenship has been considered a 'reward' for individuals who have overcome a number of obstacles and who are therefore deserving of a place in the community (Böcker & Strik, 2011; Adamo, 2008; Kostakopoulo, 2010). Previously, undertaking these steps was viewed primarily as an incentive for integration, and much less as a reward.

Hogan-Brun et al. (2009b) locate the shift in language policies for immigration in Europe around the year 2004, linking it to the expansion of the European Union to formerly communist countries which led to massive immigration from Eastern EU countries to Western EU countries. Some authors, however, attribute the increase to other economic and political factors (e. g. Van Avermaet, 2009; Saville, 2009; Hargreaves, 2010; Slade, 2010a): globalisation, general labour migration, education, tourism and political persecution. The resulting increase in multiculturalism and multilingualism in the host countries was often perceived as a threat to national identity, even in countries such as Australia which were built on immigration (Slade, 2010a; Wright, 2008). According to Hogan-Brun et al. (2009b), this effect was particularly notable in the EU member states, where the ongoing political and economic unification processes alone had led to a crisis in national identity. As a result, many nation-states reacted by issuing stricter immigration policies in the interest of preserving national identity and social cohesion. Increasing xenophobia in the course of the 2000s, further nourished by several terrorist attacks, is regarded as another factor leading to stricter immigration policies and the introduction of regulations geared to heighten 'national security' (Blackledge, 2009c; Laversuch, 2008; Van Avermaet, 2009; Wright, 2008).

2.2 The implementation of stricter immigration policies in language-related areas

Changes in language policies have given rise to an astounding variety of language requirements in different countries. Variation is found regarding test methods, test design and administration, skills tested, required level of language proficiency, test takers' minimal age, fees, exemption of certain groups, and the status of preparatory language and integration courses (compulsory *vs*. voluntary). The diversity is highlighted in a number of recent studies on testing regimes in European countries (see Van Avermaet, 2009; Böcker & Strik, 2011[4]; Extramania & Van Avermaet, 2010). On the basis of data collected from 17 countries, Van Avermaet (2009) found that the language level required can range between A1 and B1 for admission to the country, A1 and B2 for permanent residency and A2 and B2 for citizenship. These data reveal striking differences in the language requirements of individual countries. Similarly, in her comparison of practices in English-speaking countries (the USA, Canada and the UK), Hargreaves (2010) identifies significant differences between the three countries both in the format of the citizenship tests used as well as in their content. The UK and Canada, for instance, employ a multiple-choice test

4 | Böcker & Strik (2011) report only on the results concerning the specific case of application for permanent residency, while the larger original study (INTEC study, see Strik et al., 2010) addresses the different stages of the immigration process: admission to the host country, settlement and naturalisation. This study not only provides a detailed overview of the testing regimes in each of the countries analyzed; it also collected data of the impact of such language tests by means of informal interviews with different stakeholders, with the aim of identifying their point of view on the practices and their consequences.

(computer-administered in the first case, paper-and-pencil-based in the latter case), while in the USA an informal interview is conducted. According to some authors (Hargreaves, 2010; Hogan-Brun et al., 2009b; McNamara, 2009), such variation in practices and testing regimes is partly due to the fact that the testing regimes are decided upon and introduced by national governments (instead of experts) in response to the prevailing sociopolitical context and the perceived attitudes of the population. All studies provide evidence of rapidly evolving and fairly uncoordinated practices in the individual countries, especially across Europe.

The following sections further highlight measures taken to implement stricter immigration requirements with respect to the language(s) of the host country: first, the recent introduction or renewal of language requirements for immigration and integration purposes in various countries, and the extension of such requirements to more stages in the immigration process; second, the increased use of formal language tests in order to enforce requirements; third, the language proficiency level required; fourth, the knowledge-of-society tests and their implicit demands on language knowledge.

2.2.1 Proliferation of language requirements

Language requirements have become a core aspect of immigration and integration policies in many of the countries considered in this report. Not only have more countries introduced such requirements; language requirements have also been extended to more stages of the 'migrant's journey' (Saville, 2009).

Several studies report a dramatic increase in language requirements that is sometimes accompanied by higher fees in the immigration process and by sanctions for those who do not meet the requirements (Hogan-Brun et al., 2009b; Van Avermaet, 2009; Böcker & Strik, 2011). Based on data from various official reports, Extramania (2012) locates a peak in the introduction of language requirements in Western European countries (Denmark, Belgium, Germany, Greece, Norway, Austria, the Netherlands, France and Liechtenstein) between 2003 and 2008[5]. For Eastern European countries (Slovakia, Armenia, Slovenia, Poland and Turkey), such a peak can be observed between 2007 and 2009. Extramania (2012) further reports that the majority of the countries with newly introduced language requirements for immigration and integration purposes also offer language courses to the candidates.

Interestingly, once stricter language policies and testing regimes have been introduced in one country, they tend to be copied by other countries. For example, when the Dutch government introduced language requirements for permanent residency and admission to the country in 2006, five other EU member states followed the Dutch lead by adopting requirements for the same purposes the same year (Böcker & Strik, 2011; Van Avermaet, 2009).

In addition, there is a general tendency to attach language requirements progressively to all three stages of the immigration process. Originally, language requirements were introduced in most countries for naturalisation (according to Böcker & Strik, 2011) they became common practice in the 1990s). Then, they were imposed on individuals wishing to attain permanent residency. Using numbers from two surveys on language policies for immigration in several European countries commissioned by the Council of Europe, Extramania (2012) found a considerable increase between 2007 and 2009 in the proportion of countries imposing language requirements as a condition for permanent residency: while, in 2007, 57% of the countries participating in the survey had language requirements, this figure rose to 69% in 2009.

Even more recently, but in fewer cases, language requirements were established as

5 | Mar-Molinero (2006) mentions Spain as one notable exception to this general tendency. Citizenship is hard to obtain in Spain. It has a very low rate of naturalisation, no language requirements for residency and citizenship, no overt language tests, but instead an interview with a judge, in which 'adaptation' to local culture is indirectly assessed. The absence of language requirements (knowledge of the 'national' language) for citizenship is related to the country's history (avoiding dissent from linguistic minorities from the Comunidades autónomas).

a condition for admission to a country (i. e. granting a visa) for example by the Netherlands (Extra & Spotti, 2009b) and Germany (Michalowski, 2010). Extramania (2012) reports an increase from 19% of the surveyed countries applying this condition in 2007 to 26% in 2009. According to Michalowski (2010), this specific restrictive measure in both the Netherlands and Germany seems to be a response to increased family immigration. Thus, imposing a language test before arrival in the country becomes an indirect means of limiting family immigration.

Language requirements as a pre-entry condition are commented on rather critically (Kostakopoulo, 2010; Goodman, 2011) and are rooted in the argument that the integration process can only start once an individual is physically present in a country and has the chance to interact with the host community. As Goodman (2011) puts it:

> Pre-entry integration requirements mandate a degree of integration into the state while the applicant is physically and conceptually – vis-à-vis legal status – *outside* the state (Goodman, 2011: 237).

Moreover, as Extra & Spotti (2009a) and Hogan-Brun et al. (2009b) point out, access to the target language while still living abroad might be difficult (e. g. unavailability of language courses). In such cases, applicants are expected to know a language they have not had the opportunity to learn and practice, at least not in an auspicious setting.

2.2.2 Increasing use of standardised language tests

In line with the growth of language requirements in European immigration policies, the use of formal language tests has become increasingly common (Van Avermaet, 2009). All three stages – admission, residency and citizenship – are concerned. According to Van Avermaet (2009), who reports on a survey carried out in 2007 by the Association of Language Testers in Europe in 19 European countries, nine countries out of eleven with a language requirement for citizenship implemented a language test, while six out of eight countries imposing a language requirement for permanent residency also instituted a formal language test for those wishing to obtain that status. The results for admission to the country are similar: seven out of nine countries with a language requirement for admission employ a formal language test. In other words, almost all countries with language requirements as part of their immigration policies determine whether a candidate has acquired the targeted proficiency level by means of a formal test.

In some cases, a standardised test substituted earlier forms of assessment (Hogan-Brun et al., 2009b). Böcker & Strik (2011) cite cases of former naturalisation procedures in the Netherlands and in Denmark, in which language skills were evaluated informally in an interview with a civil servant. Then, in the early 2000s, both countries introduced standardised tests to assess the language knowledge of applicants for citizenship.

2.2.3 Levels of language proficiency required

The levels of language proficiency required for admission to a country, residency or citizenship vary greatly from country to country. The overview of testing regimes in 17 European countries in Van Avermaet (2009) shows that the language proficiency levels required for *admission* vary between A1.1 and B1 on the Council of Europe scale, for *residency* they ranged from A1 to B2, and for *citizenship* from A2 to B2. Referring to Germany, where level B1 is required for citizenship, Möllering (2010) observes that this relatively high level of proficiency might discourage people from applying for naturalisation, and suggests that level A2, as required in the Netherlands, might be more suitable.

The language proficiency requirements are not always described with reference to a defined framework of levels, such as the *Reference Levels* in the *Common European Framework of Reference for Languages* (CEFR – Council of Europe, 2001). Canadian legislation, for example, requires 'the ability to speak English or French well enough to communicate with people' (Hargreaves, 2010); the Australian Citizenship Test requires 'a basic knowledge of English' (Farrell, 2010; Möllering & Silaghi, 2010). In this latter case, the Woolcott Report (see Möllering & Silaghi, 2010); McNamara & Ryan (2011) criticised the vagueness of this language proficiency description – that it remained unclear as to what exactly needs to be known – and reprimanded its lack of transparency, as this makes it difficult for applicants to know how to meet the conditions for a successful application.

Böcker & Strik (2011) draw attention to the fact that the proficiency levels required for residency and citizenship tend to rise when language requirements are defined even as a pre-entry condition. In other words, the introduction of a language requirement at an early stage of the immigration and integration process apparently raises the bar for the latter stages.

2.2.4 Language testing and cultural knowledge testing

In a number of countries, such as the UK, Denmark or the Netherlands, applicants for residency and citizenship must take a test of cultural and society knowledge in addition to the language test (Böcker & Strik, 2011; Van Avermaet, 2009). In culture and society knowledge tests, the candidates are evaluated on topics such as the religion, culture, law and history of the host society. In tasks like knowing and understanding the Australian Pledge of Allegiance (Slade, 2010a), Möllering (2010) identifies an assimilationist approach to citizenship. According to Slade (2010a), the rationale for the use of such a test is typically based on the assumption that, by acquiring knowledge of sociocultural facts, norms and values, the applicants will actually adhere to the values of the society they learned about – a belief that, in her opinion, can hardly be confirmed.

Two additional problems regarding knowledge-of-society tests are brought up in the literature analysed. First, although they seem to address plain, factual knowledge, some questions actually address the candidates' attitudes and personal beliefs (e. g. on homosexuality or religion). This is judged as inequitable (Joppke, 2010; Michalowski, 2011), and a potential source of social discrimination. Second, since knowledge-of-society tests are often taken in the official language – or one thereof – of the host society, success depends on the candidate's language skills. McNamara (2009b) categorises tests of knowledge of society as *de facto* language tests. Extra & Spotti (2009a) point to the Dutch language-and-culture tests as examples of this. McNamara & Ryan (2011) provide details on the controversy accompanying the introduction of the new Australian Citizenship Test; objections were raised because the language knowledge necessary to complete the test exceeded (and, even after revision, still exceeds) the officially required level for citizenship.

In sum, the studies referred to in chapter 2 take a critical stance on recent developments concerning language requirements for immigration. They provide evidence of a tendency in the Western countries, Europe in particular, over the past ten years to introduce stricter language proficiency requirements at the different stages of immigration. Furthermore, a strong trend from informal to formal language testing is observed. While these overall tendencies emerge rather clearly, the studies reveal considerable diversity between countries regarding the language proficiency levels required. Cultural knowledge tests are generally criticised in the studies, not only because of the type of knowledge they refer to, but also because their implicit language requirements exceed the language proficiency levels defined.

3

Validity issues in language testing for immigration and integration purposes

This chapter presents how and to what extent the literature on language testing for immigration and integration purposes discusses aspects of assessment validity. The quality and suitability of (formal) assessments are usually established by means of a validity or validation framework that guides confirmatory or critical discussions and analyses. The following presentation of validity issues is organised along categories taken from a validation framework that is both comprehensive and widely accepted within the language testing profession: Bachman and Palmer's framework for assessment validation and validity inquiry, the so-called Assessment Use Argument (AUA). In a first step, the essentials of this framework are briefly introduced. Subsequently, topics from the literature relating to validity are highlighted using categories from the AUA framework, then complemented by some necessary additions.

3.1 Bachman and Palmer's argument-based approach to assessment validation

The validity of assessments has been a recurrent topic of central importance in psychological, educational and language testing for several decades. Bachman and Palmer's Assessment Use Argument (AUA) framework was recently described and operationalised in great detail in (Bachman & Palmer, 2010). It builds on two main sources: a) conceptual and practical work previously carried out by various authors[6], including authoritative associations in the field[7]; b) Bachman's and Bachman & Palmer's earlier work on task-based communicative language testing (Bachman, 1990; Bachman & Palmer, 1996; Bachman, 2005). Chapter 5 of Bachman & Palmer (2010) provides a concise overview of the AUA framework.

The main motive for the use of a validation framework like the AUA lies in the fact that (summative) language assessments are typically used for decisions that entail consequences for stakeholders, such as the candidates themselves, teachers, employers or society at large. Since individuals are affected by assessment-based decisions, Bachman & Palmer see an obligation[8] for test developers and test users (e. g. immigration authorities) to be accountable to these individuals. Being accountable means that they 'must *demonstrate*, through argumentation and supporting evidence, that the use of a particular assessment is justified' (B&P[9]: 85-86). Once established, this argumentation consists of a series of well-supported claims to validity that are interconnected (through inferences) to form a comprehensive validity argument for a specific use of an assessment. The usual quality criteria for (formal) assessments, such as test and (inter-) rater reliability, lack of group bias or content validity are integral parts of these claims and the validity argument as a whole (see our references to these criteria in square brackets in the list of

6 | Most notably Messick (1989); Messick (1994); Kane et al. (1999); Kane (2004); Kane (2006); Mislevy et al. (2002); Mislevy et al. (2003); Kunnan (2004).

7 | 01 Standards: American Educational Research Association, American Psychological Association, National Council on Measurement in Education & Joint Committee on Standards for Educational and Psychological Testing (U.S.) (1999); Educational Testing Service (ETS) (2002); Codes of Practice: ILTA (International Language Testing Association) (2007); ALTE [Association of Language Testers in Europe] (1994); ALTE [Association of Language Testers in Europe] (2001); handbooks: Linn, National Council on Measurement in Education & American Council on Education (1989); Brennan, National Council on Measurement in Education & American Council on Education (2006)

8 | Bachman & Palmer introduce the ***need to be accountable*** to the individuals concerned and the ***obligation to demonstrate*** that the use of an assessment is justified as 'two foundational axioms' of their approach (Bachman & Palmer, 2010: 85).

9 | short for Bachman & Palmer (2010)

claims below). The main task that falls to test developers and/or users is to demonstrate that the various claims are met to a sufficient degree, given the assessment's purpose and consequences. A valid AUA equally justifies a specific test method and the use of a test for a specific purpose in society.

In the following, the various claims that should be supported through an AUA are briefly listed.

Claim 1 – intended consequences
The *consequences* of using an assessment and of the decisions that are taken are **beneficial** to stakeholders (test takers, teachers, and society at large) (B&P: 105).

Claim 2 – decisions
The *decisions* taken on the basis of the interpretations of test scores…

→ take into consideration community **values** and relevant legal requirements, and
→ are **equitable** for those stakeholders who are affected by the decision [lack of bias] (B&P: 111).

Claim 3 – interpretations
The *interpretations* about the ability to be assessed are…

→ **meaningful** with respect to a particular learning syllabus, an analysis of the abilities needed to perform tasks in the TLU[10] domain, a general theory of language ability or any combination of these [construct validity],
→ **impartial** to all groups of test takers [lack of bias],
→ **generalizable** to the TLU domain in which the decision is to be made [construct validity],
→ **relevant** to the decision to be made [construct validity], and
→ **sufficient** for the decision to be made [construct validity] (B&P: 114).

10 | TLU: Target-Language Use, i. e. 'real-world' communicative language use.

Claim 4 – assessment records
The *assessment records* (scores, descriptions of test-taker performances) are **consistent** across different assessment tasks, different aspects of the assessment procedure (e. g. test forms, occasions, raters) and across different groups of test takers (B&P: 124) [reliability; lack of bias].

As mentioned earlier, comprehensive test validation and the provision of convincing validity evidence are highly desirable in any high-stakes context such as decisions on entry, residency and citizenship. Despite this, an overview of the literature on language assessment for immigration and integration purposes covered in this paper reveals immediately that a comprehensive validity argument is rarely even attempted. Many papers focus on a few selected aspects of validity only, with issues related to claims 1 (*beneficence*), 2 (*values and equitability*) and 3 (*meaningfulness*) being made the object of discussion more frequently than other aspects. These discussions are, however, often quite general, i. e. not related to a specific assessment and its use.

Only very few publications are actually geared towards providing (and/or probing) a comprehensive validity argument, or systematically analysing a series of critical features of a single test and its specific use(s). This is unsurprising, as publications focusing on only one test from the field of testing for immigration and integration purposes are generally quite rare.

The purpose of the following sections is to reconstruct the discussions in the literature on language testing for immigration and integration purposes in relation to each of the main categories (i. e. claims to validity) used in the Assessment Use Argument. This means that it is not the validity arguments for individual assessments that are investigated here, but rather the issues raised in general and across contexts as they pertain to each of the main aspects of validity dealt with in the AUA.

In addition, the AUA's system of categories must be adapted in order to accommodate those studies that discuss language testing for immigration and integration purposes from a more radical point of view and that raise the question whether language testing – for various reasons – is *just*, i. e. a legitimate means in the immigration context *at all*.

McNamara and colleagues (McNamara, 2006; McNamara & Roever, 2006; McNamara & Ryan, 2011) have made substantial contributions to the discussion of whether assessment use is 'just'. They see themselves in the tradition of Critical Language Testing (Shohamy, 2001; Shohamy, 2006), which focuses on implicit and explicit relationships between language and power. At the same time, they build on Messick's seminal publications on validity (Messick, 1989; Messick, 1996; McNamara, 2006), which emphasise the dimension of *assessment use* as an integral part of validity.

In a recent publication on the topic of *justice* in language assessment, McNamara & Ryan (2011) present 'fairness' and 'justice' as two complementary concepts that are equally necessary for validation in (language) testing. Questions of test *fairness*, as they define it, involve not only a concern with equal treatment of groups and the avoidance of psychometric bias, but include all aspects of the empirical validation of test score inferences in the interest of reasonable and defensible assessments of individual test takers. Questions concerning the *justice* of tests complement this approach by additionally considering the consequences of test score interpretation and use (or misuse) in a specific social context[11] as well as the social and political values *implicit* in test constructs.

Arguably, all of 'fairness' and some aspects of 'justice' are covered by Bachman and Palmer's Assessment *Use* Argument. Bachman (2005) and Bachman & Palmer (2010) illustrate briefly how misuses of tests brought to the fore by Critical Language Testing (Shohamy, 2001, in particular) could be accommodated by the AUA. They argue that many potential misuses mentioned could be counteracted by establishing the quality requirement 'beneficence of consequences' for the various stakeholders.

In their discussion of Messick (1989), however, McNamara and Ryan (2011) disagree with this argument and point out that language testing cannot be self-sufficient in dealing with 'issues of social value'. According to them, testing must also

> rely on other kinds of analysis, more familiar in cultural analysis and critical policy analysis, often considering questions of history, ideology, and discourse context and using primarily qualitative analytical tools (McNamara & Ryan, 2011).

3.2 Justice and language assessment for immigration and integration purposes

This section provides an account of the discussion on 'issues of social value' (McNamara & Ryan, 2011) in the literature on language assessment for immigration and integration purposes. It focuses on two areas criticised by critical language testers that are partially related with each other:

1. Social and political values and ideologies implicit in test constructs;
2. Social gate-keeping by means of language assessments that is often based on a 'hidden agenda'.

A considerable portion of all studies available to us identify and criticise ideologies and unquestioned values that are used as a basis and justification for introducing language (and cultural) knowledge requirements and their formal assessment.

One such ideology is the myth of language as a symbol of national identity and of belonging to the community (Shohamy, 2009; Extra et al., 2009; Van Avermaet, 2012; Piller, 2001). Milani (2008), who analyses texts and discourses on the topic of immigration in the public sphere in Sweden, identifies the recurrent belief that lack of knowledge in the Swedish language is the main cause for immigrants' low engagement in political, social and economic life. Proficiency in the national language is hence perceived as a necessary condition for 'good functioning' in the host society and also as a condition for understanding the cultural dimensions of the

11 | McNamara & Ryan (2011) mention Shohamy's 'hidden agendas' – language tests covertly serving social policy – as an issue of justice.

community. According to Milani, there is an assumption

> that knowledge of a common language (Swedish) does not merely give access to the civic domain of rights, duties, and political participation and the economic sphere of the labor market, but is actually the ONLY way that immigrants will properly understand a given society, together with its laws, life, and cultural norms (Milani, 2008: 44).

Mastery of the national language is therefore taken as an indicator of successful integration in the host community, while the lack of proficiency in that language is often interpreted as the immigrant's non-willingness to integrate (Shohamy, 2009). Some interpretations go even a step further by concluding that a lack of proficiency (by inability or by refusal) in the national language is a threat to social cohesion and national identity (Blackledge, 2009c).

Some authors such as Oers (2010) and Mackenzie (2010) point out that, parallel to the rise in requirements for residency and/or citizenship, the concept of citizenship itself has changed, too[12]. Mackenzie (2010) notes a shift from a concept of citizenship as a passive phenomenon, based on a set of rights and duties, to citizenship as an active phenomenon that implies commitment to the prevailing political institutions. As such, citizenship is taken as an indicator of individual community membership:

> Citizenship is viewed as a sign that SYMBOLIZES the institutional recognition of the fulfilment of a set of prerequisites that are allegedly indispensable if an individual is to be awarded the identity of Swedish citizen, and thereby fully enter Swedish society (Milani, 2008: 43).

In this understanding, there is a one-to-one relationship between successful integration and the acquisition of citizenship. This, however, is obviously a simplification, as it has been shown that some individuals are integrated in society without being naturalised (Farrell, 2010), while others who have been citizens their entire life are badly integrated. Citizenship is also considered as a reward, something that is valuable insofar as it entails not only rights and obligations, but also privileges (Cox, 2010).

A concept of citizenship as the culminating point of the integration process through which an individual proves his or her commitment to the host society is closely related to the notion of citizenship as assimilation: the prospective citizen is expected to abandon his or her previous identity and values and to fully adhere to the cultural and social norms of the host society (Kostakopoulo, 2010; Slade, 2010a).

According to some authors, the nationalist ideology of 'one nation – one language' comes into play in most European countries that require a specific level of proficiency in the 'national' language – to the detriment of other regional varieties (Extra & Spotti, 2009a; Blackledge, 2009a; Shohamy, 2009) – thus consolidating the existing linguistic and cultural hegemony (Hogan-Brun et al., 2009b; Shohamy, 2009). Immigrants are expected to demonstrate their integration in the host society by learning the standard language and adhering to that country's cultural values. Blackledge (2009c) goes as far as to say that language testing regimes, by setting standardised levels of English as requirements, even have to invent or construct 'English' as a homogeneous set of linguistic practices while, in actual fact, there is no such thing as standard English because practices change across contexts. Van Avermaet (2009) draws attention to the fact that language and cultural knowledge testing are generally based on the assumption that societies are homogeneous in terms of language and cultural norms because simplified constructs lend themselves to testing.

For McNamara it is clearly the symbolic function of knowing the national language that stands in the foreground:

> The motivation for the inclusion of a language requirement is not primarily about the communicative but about the symbolic function of language. The primary function of the test is not to promote the

12 | According to Oers (2010), three concepts of citizenship can be found in the literature: a) citizenship as a legal status (a privileged relationship between a person and a state), b) citizenship as an activity (participation in the social life of the polity), and c) citizenship as identity (membership in a polity, identification, personal loyalty, commitment to the culture of the society).

> welfare of immigrants, but to express an ideology associating language use with cultural values (McNamara, 2009: 158).

Consequently, McNamara argues for a reformulation of the test construct, as these tests do not actually measure functional language proficiency (e. g. practical communicative skills that might be relevant for the test takers in their everyday lives as residents/citizens), but rather implicitly assess 'conformity to a set of socio-cultural values', or a 'national ideology' (McNamara, 2009b). In actual fact, they are scored on the basis of the dichotomy 'acceptable' *vs*. 'inacceptable' – established by policy makers rather than language testers – and interpreted as eligible, or ineligible, for residency/citizenship (McNamara, 2010).

For Kostakopoulo (2010) the link between knowledge and skills testing and adherence to the host society is not obvious, either:

> Indeed, it is very rare to find an answer to the question of why it is presumed that 'shared belonging' is something that can be obtained by testing one's fluency in the host language and the accumulation of factual information about civics, history or life in the country, which may well be forgotten a few months after the test, rather than on the basis of shared common experiences, working and contributing to the common good and enhancing the welfare of the society (Kostakopoulo, 2010: 9).

From this point of view, the usual testing that is carried out does not seem to be very relevant.

For Milani (2008), language testing in the context of immigration hardly promotes integration, but does rather the opposite: it promotes social dissociation. He believes a contradiction arises between the overt aim of tests for immigration and integration purposes as advertised in public discourse and the actual consequences of these tests. While language tests for citizenship are supposed to support the inclusion of immigrants in the host society – because social cohesion and national identity are believed to be achieved only through a common language –, the effect of these language tests in reality is rather exclusion insofar as some people inevitably will fail the test. Milani (2008) shows, on the basis of interviews carried out with stakeholders, that immigrants themselves perceive these language policies as being designed for the purpose of excluding certain groups of people. Milani (2008) points out that language tests contribute to social dissociation in that they draw a new boundary between social categories: they oppose the group of those who succeed in passing the test (and hence acquire citizenship), to those who fail (and retain their position as non-citizens). They also add to the distinction between the group of immigrants who wish to become citizens and are therefore obliged to pass a test, and the nationals who are not required to pass this test. Therefore, instead of reducing social differentiation as it is officially claimed, language tests for naturalisation contribute to reproducing difference by excluding certain groups of people from a set of civic and cultural domains (Milani, 2008).

Authors arguing from a Critical Language Testing (Shohamy, 1998; Shohamy, 2001) stance contend that language assessments for immigration and integration purposes are (mis)used by governments as instruments for *gate-keeping* (Shohamy, 2009; Blackledge, 2009c; Hogan-Brun et al., 2009b; Van Avermaet, 2009; Slade, 2010a; Wright, 2008) and for preserving the privileges of the nationals (Piller, 2001). Using language assessments for this purpose allows governments not only to control the flow of immigrants, but also to demonstrate to the population that everything is under control (Böcker & Strik, 2011; Slade, 2010a). The political actors can claim they have taken the steps necessary to secure social cohesion (McNamara, 2009b) by preserving the nation from too much linguistic and cultural diversity (Kostakopoulo, 2010; Milani, 2008).

In social gate-keeping, language is taken as an indicator of membership in a social group and the non-compliance with language requirements leads to the denial of citizenship and, consequently, to social exclusion. Milani (2008: 53) points out the tendency of some nation-states to 'raise the MEMBERSHIP BAR that regulates access to the in-group' by introducing language assessments whenever faced with increased immigration flows. The old Australian Dictation Test (McNamara, 2005) is quoted as an extreme case of misuse of a language test by a government. Similar to the Shibboleth test in biblical times, the Dictation Test deliberately aimed at discriminating less welcome social groups by ordering prospective immigrants to pass a test in a language they did not know, thus undermining any chance for success (McNamara, 2005).

Siiner (2006) reports on a recent case of gate-keeping in Estonia that has had serious or even detrimental impact on a great number of individuals. With the aim of rebuilding the nation after the end of the Soviet occupation and imposing Estonian as the only national language, the government introduced a requirement to know Estonian, and tested basic conversational and writing skills as a condition for obtaining national citizenship. Many Russian-speaking residents were not able to fulfil this requirement, as they were not proficient enough in the national language. These individuals were withheld citizenship even if they were born in Estonia, and even if, in many places, people could function very well in everyday life using Russian only. In their position as non-citizens, Russian speakers were deprived of basic rights; for example, they were not allowed to apply for certain positions in the public sector. As a result, a large portion of the Russian-speaking minority remained stateless, as they were neither citizens of Estonia nor of Russia, and ended up being marginalised from society. The exclusion of minority groups in the Baltic states on grounds of insufficient language competence in the national language, legitimated by the national language policies, has been largely criticised by European institutions (Ozolins, 2003).

Shohamy discusses the question whether language assessments for residency and citizenship are just from a more radical stance, by considering ethical and fundamental political arguments. Shohamy (2006) expresses her opposition to language requirements and language assessment for immigration and integration purposes. In her view, it is unethical that individuals who do not know the language that is dominant in a given context are deprived the opportunity to fully participate in society as citizens. Shohamy further argues that, since this deprivation is unethical, the refusal to grant citizenship to someone because of insufficient language proficiency is a violation of his or her personal rights – a discriminatory practice. Civic rights and obligations are granted to individuals who master the dominant language, while these same rights and obligations are denied to those who do not. Shohamy argues that the refusal to grant rights on the basis of deficient language proficiency is not a fair motive because 'there is no indication that being proficient in the national language necessarily creates better citizens' (Shohamy, 2006: 148). Moreover, Shohamy (2009; see also Blackledge, 2009a) questions a state's legitimacy to impose on any individual learning and using a given language through language policy and language testing. She defends her position by stating:

> One wonders whether the acquisition of 'national' languages should not be a choice for people who can make their own best rational decisions as to the language they need to know and use in a multilingual, transnational world (Shohamy, 2009: 49).

In other words, obligating people to learn a national language constitutes a violation of their right of freedom to speak the language that is best fitted to their daily needs; it furthermore transforms knowledge of a national language to a 'civic duty' (Shohamy, 2009).

In light of the above, Shohamy and other critical scholars (e. g. Blackledge, 2009c; Cox, 2010; McNamara & Shohamy, 2008; Van Avermaet, 2012) denounce the social consequences of a language requirement for accessing residency and citizenship as well as the political instrumentalisation of language tests for that purpose. Through their writings, they try to raise the awareness of test users for the impact these tests have on the actual test takers' lives and call upon them to take responsibility for the (mis)uses of these tests.

Overall, the notion of justice is used as a device for – often radical – criticism of language requirements and language testing in the context of immigration. Ideological thinking, such as 'one nation – one (standard) language', is revealed behind language requirements that seem to be set for the test takers' own good, i. e. integration. Also, according to the critics, language and cultural knowledge tests primarily assess conformity to a set of socio-cultural values, as there is no evidence for the claim that the constructs tested are crucial for social participation. In addition, these scholars point out that, in reality, language requirements and testing result in even greater social dissociation and exclusion (testing as gate-keeping) rather than integration, which authorities often declare to be the goal. Finally, there is a strand of criticism that holds language requirements and their consequences to be an infringement on an individual's fundamental rights and, therefore, as an unethical and discriminatory practice.

3.3 The beneficence of assessment use

According to the AUA conceptual framework, it is a sign of assessment validity when the consequences of assessment use and decisions are beneficial to all stakeholders and not detrimental to any. This relates to individuals as well as society at large.

Some arguments related to the beneficence of assessments were discussed in the section on justice above, but from a general (and fundamentally critical) perspective only. In this section, the focus is on concrete rather than general aspects reported in studies dealing with the impact of language requirements on individuals (e. g. access to employment or education) and society (e. g. effects on social cohesion). A first subsection summarises points from a variety of publications on putative or actual effects of language requirements and testing regimes. A second subsection focuses specifically on findings from dedicated impact studies.

Typically, tightened conditions for immigration and integration purposes are presented by policy-makers, and sometimes test developers, as being motivated by practical considerations, and the potentially beneficial effects on the immigrants are highlighted (Blackledge, 2009c). De Jong et al. (2009) make an obvious effort to justify the use of their new Test of Spoken Dutch by highlighting the beneficial effects of better language proficiency for test takers and society. According to them, lack of integration (due to language deficits) leads to 'a vicious cycle, in which parents cannot help their children, children drop out of schools, leading to feelings of hopelessness and despair', while testing helps to fight the 'negative consequences of the social segregation of large numbers of immigrants in the Netherlands'. Positive effects are also asserted with regard to testing before arrival in the host country; these tests are said to help prevent forced marriages and female trafficking, 'a new form of slave-trade'. As an additional argument, the results of a survey conducted before actual test development started are cited. According to the developers it showed that

> despite its potential for political controversy, the purpose and intention of the law on the integration for new residents were supported and considered justifiable by a large majority both inside and outside parliament [even by immigrants interviewed] (De Jong et al., 2009: 43).

The spirit of the arguments brought forward by these authors can be found elsewhere as well. The usual rationale for the introduction of language requirements in immigration policies is that they facilitate an immigrant's integration in the host community and promote his or her active participation in everyday and civic activities (Blackledge, 2009c; Oers et al., 2010; Holland, 2010), thus enabling them to be economically independent (Böcker & Strik, 2011). In view of these benefits, all immigrants should be equipped with a sufficient degree of 'functional' language competence (Blackledge, 2009c). For example, the prescription of level B1 according to the Common European Framework of Reference for Languages (Council of Europe, 2001) for immigrants in Germany wishing to acquire permanent residency or citizenship was motivated by the aim of making immigrants independent in both the private and the professional sphere (Oers et al., 2010). Introducing a fairly demanding language test is therefore justified as a way to guarantee economically self-sufficient and socially active new citizens. Additional courses and tests of knowledge about history, culture and laws of the host society serve to complement the repertoire necessary to fully participate in social life (Cooke, 2009; Milani, 2008).

As demonstrated in the section on justice above, a rationale that resorts to the beneficence of the prescribed measures is viewed very sceptically in a number of studies. There are, however, other studies that point to the positive effects of (stricter) language requirements as fact. Kiwan (2008), for example, argues that preparing for the British citizenship test can be seen as an opportunity for immigrants to acquire the relevant knowledge about their rights and obligations as future citizens. In that sense, the test can be seen as a tool promoting the understanding of what active civic participation means. It is noteworthy that it is not the language test itself that is credited for contributing to integration, but rather the preparation in the context of an integration course. Similarly, Yoffe (2010) sees positive effects in the Reception and Integration Contract that was introduced in France in 2007. This contract obliges newly arrived immigrants to attend language training if their proficiency level in French is lower than A1.1 according to the CEFR. For Yoffe, the obligation to attend language courses offers an opportunity – notably to uneducated Muslim women – to leave their prescribed social environment and communicate with native speakers; in the absence of an obligatory course, these women would not be allowed to socialise with local people. Another positive effect Yoffe observes is that immigrants who successfully complete their language training gain 'confidence in their ability to function in the society and the motivation to continue language study' (Yoffe, 2010: 77).

Empirical studies investigating the consequences of assessment use are traditionally known in the field of validation research as *impact studies (*or *washback studies* for educational contexts). These studies identify the consequences of language tests on society at large (e. g. effects on social cohesion) and on the lives of the test takers (e. g. better access to employment or education).

Impact studies on language assessment for immigration and integration purposes are rather rare, partly due to the fact that the introduction of formal requirements and tests is a recent phenomenon (Böcker & Strik, 2011). Van Avermaet & Rocca (2013) observe that impact studies tend to be limited to statistical accounts, such as test passing rates, or the number of participants in preparatory courses. Therefore, they yield little information about the impact of a test regime on integration processes or social participation. Consequently, the two authors call for more research on the social impact of integration policies on immigrants' integration.

One important recent study on just that kind of research is the INTEC Study (Strik et al., 2010). It examines the actual practices resulting from language policies and the use of tests at the different stages of immigration (entry into country, application for permanent residency and citizenship) as well as the effects of the respective testing regimes in nine EU member states (Austria, Belgium, Denmark, France, Germany, Hungary, Latvia, the Netherlands, and the United Kingdom). The study provides two types of data: a) statistical data on the number of permits granted and refused; and b) qualitative data from a total of 329 interviews with different stakeholders (mainly immigrants required to fulfil integration requirements, but also language teachers, officials and NGO staff).

For some countries, *statistical data* on naturalisation procedures reveals a decline in the number of immigrants obtaining citizenship since the conditions were tightened. In Denmark, for example, the number of refusals to grant citizenship increased in 2002-2003 when formalised tests were introduced, and again in 2007-2009 when the required language proficiency level was raised from B1 to B2. Similarly, Oers (2008) notes a significant decrease in applications for Dutch citizenship in 2004 (70% less applications in 2004 compared to 2002) immediately after the introduction of the naturalisation test. In the case of the two countries that introduced tests to be taken before entering the country (the Netherland and Germany), the INTEC study established a drop in the number of applications, which essentially concerns family reunifications. This means that for at least some and probably most of the women who wished to join their partner (either husband or future husband) the new policy had non-beneficial effects. If a test was failed, families could not be reunited and marriages could not be concluded.

According to the INTEC Study, the *qualitative data* also indicate that the immigrants themselves often perceive the effects of the new policies as non-beneficial. As far as the naturalisation procedures are concerned, immigrants to Austria, Denmark, Germany, the UK and Latvia were sceptical of a future, positive impact of these tests on their integration process – in the domains of employment and social life, for example. Rather, the respondents attributed successful integration to other factors. This is not very surprising as most of them already felt integrated by the time they applied for naturalisation. As negative effects, some immigrants explicitly pointed to the stress and anxiety they (and their families) experienced when they had to pass a test whose outcome would have a considerable impact on their life. The preparatory courses were criticised as well for being time-consuming and incompatible with work obligations.

The INTEC Study (Strik et al., 2010) also reports some positive, i. e. beneficial effects. For example, German-language teachers observed cases where the applicants intentionally failed the pre-entry test in order to avoid a forced marriage. Some respondents to the study who applied for permanent residency agreed in principle that the preparatory course they took had an 'emancipatory effect'. Compulsory language courses (in Germany or Austria, for example) apparently constitute a unique opportunity for some women to socialise because they would not be allowed to attend an integration programme on a voluntary basis. The possibility to have a social life has demonstrably positive effects on the self-esteem of these women, who generally have had little formal education.

The results of a smaller social impact study in Flanders, presented in Van Avermaet & Rocca (2013) and Van Avermaet (2012) show little positive impact of the testing regimes. The study is based on 40 semi-structured interviews with various stakeholders (immigrants, teachers and representatives of the 'majority' group, for example, employers). In Flanders, newcomers and already settled immigrants are required to attend compulsory integration courses to reach level A1 in Flemish and to acquire basic knowledge of society. At the end of the integration course, successful participants receive a certificate of integration. The teacher makes the decision on whether the course was passed or failed.

Interviewed immigrants who attended the course more than a year earlier and who did not find a job following the course unsurprisingly stated that the integration course and the certificate were not useful for gaining access to employment. Interviews with employers and job agencies revealed that the certificate of integration is not even taken into consideration in job applications and that the immigrants' language skills are informally assessed in the job interviews. Generally speaking, language competence does not emerge as an important factor for access to the job market in Flanders. The study concludes that, overall, integration measures have a rather limited impact on an immigrant's actual integration process.

In sum, the studies available demonstrate that the purported beneficence of language requirements and language tests for immigration remains controversial. Policy-makers and test developers alike have a tendency to legitimate new regulations by highlighting potential benefits for immigrants and society: better integration and more economic independence through better knowledge of language and society. But, the few available studies support these claims only marginally: some scholars observed immigrants who benefitted from new testing regimes through the preparatory courses, among them Muslim women who had an opportunity to socialise outside their families. Impact studies based on statistical data and interviews with an adequate selection and number of stakeholders show mostly negative or no consequences: where language tests are mandatory in order to get a visa, fewer individuals (often women wishing to get married) can immigrate; where a language certificate is necessary to apply for naturalisation, fewer residents obtain citizenship; for applicants who are already well integrated, an integration course and exam primarily mean a packed schedule and more stress. One study shows that the certificates issued for successful completion of the requirements have little or no value on the job market. Therefore, overall, there is little evidence in the literature to back optimistic claims about the beneficence of stricter requirements regarding language proficiency and cultural knowledge. It is, however, also clear that additional impact studies are much needed.

3.4 Values-sensitivity of decisions

Bachman & Palmer (2010) suggest that, for an assessment to be value sensitive, test developers need to engage with the values different stakeholders in an assessment may have and acknowledge them in one way or another. Going one step further, they encourage test developers to question existing community values if these put fair and equitable treatment of all candidates at risk.

Based on the literature included in this study, it can generally be said that values receive little consideration by actual test developers, while they seem to be a priority among socio-critical academics writing about language requirements and assessment.

The way Plassmann, representing telc, one of the developers of the German Test for Immigrants (DTZ), deals with the fact that candidates have to demonstrate level B1 (generally considered as demanding) at the end of their integration course, may provide an explanation for this apparent disinterest. She states simply that level B1 was set by legislation and that, therefore, a discussion about what level would be appropriate for what purpose is unnecessary (Plassmann, 2011). The Dutch developers of the Test of Spoken Dutch (TGN) are in a similar situation as the German developers – the TGN is also a test commissioned by the authorities – but they do engage, though briefly, with value issues before moving on to present the technical validation work (De Jong et al., 2009). Thus, they mention several facts and findings in order to demonstrate that their test is on firm ground with regard to values and acceptance. They emphasise that the introduction of the new test is based on a parliamentary vote with broad support, that it was commissioned by the Ministry of Justice and that this same body also selected the CEFR scale as the reporting scale. They furthermore report that the political opposition received no backing for their views, not even among immigrants. Then, they continue by explicating the expected social benefits of their test: the TGN promotes integration, with lack of integration possibly leading to a 'vicious cycle' causing hopelessness and despair; testing abroad can help control female trafficking; and those who make a minimal effort can immigrate because the test is not too demanding (De Jong et al., 2009). The developers of the Dutch test display a clearly affirmative attitude towards their test and do not appreciatively engage with opposing views[13].

3.5 Equitability of decisions

According to Bachman & Palmer (2010), whether an assessment-based decision is equitable or not depends mainly on two aspects: equal opportunity for different test takers at the same level of ability to be classified in the same group (e. g. 'pass' or 'fail'); and equal opportunity for different test taker groups to master the skills required. Membership in a particular social group should have no influence whatsoever on decisions taken in connection with tests.

The idea that formal tests are (at least potentially) more equitable than, for example, an informal interview with a civil servant is mentioned in several of our studies. Joppke (2010; similarly Wright, 2008) considers the introduction of standardised tests for naturalisation as fairer and more liberal than the prior informal assessment insofar as all candidates are treated equally. Several studies cite examples of informal test regimes that lead to an obvious lack of equitability; for example, Schneider et al. (2006) describe local naturalisation practices

13 | There is even a tinge of triumph and ridicule in some passages.

in Switzerland and Piller (2001) relates similar practices in Germany that are defined to a large degree by the preferences of individual civil servants. Laversuch (2008) shows how this situation was misused for gate-keeping purposes, particularly with regard to Muslims. The otherwise test-sceptical ALTE LAMI group mentions the advantages which 'properly designed, constructed and administered' tests have, including that 'results are highly standardised and reliable' and that 'candidates are assessed with a high degree of independence and objectivity'. They therefore advocate test development and use in adherence to high quality standards – if tests are to be used in immigration decisions at all (Balch et al., 2008).

While the potential of tests to introduce more equitability is acknowledged by various authors, highly formalised test regimes are criticised as non-equitable because they create new hurdles for certain groups of immigrants. Shohamy points out that many immigrants who are illiterate or poorly educated are unfamiliar with the testing procedures and that they rarely succeed on tests. Böcker & Strik (2011) argue that poorer people are not able to pay the required fees for taking the test. For Michalowski (2010) and Shohamy (2009), language tests are selective tools because only highly qualified and educated groups are able to surpass all obstacles and are therefore granted the desired residency status.

According to several authors, non-equitable treatment starts long before the test itself because different candidates have unequal access to information, language instruction and test preparation. This particularly concerns those settings in which a test has to be passed as a pre-entry requirement (Germany, Netherlands). Respondents to the INTEC survey (Strik et al., 2010) in Morocco draw attention to the fact that getting hold of preparation material involves travelling to the embassy and that language courses are not available where many prospective applicants live.

Shohamy (2009) observes that it may even be difficult for immigrants to come into contact with the language when living in their host country and that they often lack true opportunities to master the target language at the level required for the citizenship test.

For Krumm (2007), it is doubtful that submitting all immigrants to the same test leads to more equitability. He observes a great heterogeneity in the immigrant population in terms of linguistic, educational and cultural backgrounds, as well as differing degrees to which language skills are required to function in society and the working world. According to Krumm this poses a problem when it comes to testing these people:

> Testing people who are unequal in all aspects of their linguistic and cultural abilities and competences with one and the same test and also expecting the same level of proficiency from them in all areas cannot possibly be a way to demonstrate equality in society (Krumm, 2007: 668).

Based on statistical data on the German Test for Immigrants (DTZ), (Klein, 2013) discovered that some groups of candidates had more problems passing the DTZ at the required level than others. She therefore suggests that more support in exam preparation be provided for these disadvantaged groups: women with L1 Turkish; men with L1 Russian or Polish, as well as elderly people, particularly those with L1 Italian, Chinese or Turkish. Oers (2008) also observes that certain groups are specifically handicapped by the testing format. He mentions elderly as well as poorly or un-educated people. Oers maintains that the testing of Surinamese immigrants to the Netherlands is another example of non-equitable treatment: their first language is Dutch, but they still have to pass the Dutch test in order to attain citizenship, while there is obviously no such test for the Dutch themselves. Stevenson & Schanze (2009) locate an additional bias in the fact that many highly educated immigrants, such as EU citizens or citizens of countries like the USA or Australia, are not subjected to a language requirement and corresponding test, no matter what their knowledge of the local language is.

Overall, a mixed picture emerges on the equitability of tests. While it is recognised that tests often replaced obviously biased informal practices, some authors are careful to point out discriminatory aspects of the new testing regimes.

3.6
Meaningfulness of interpretations with respect to a construct

In order for interpretations of assessment records (i.e. interpretations of the logged test results) to be meaningful with regard to language use in a real-world context (target-language use – TLU), Bachman & Palmer (2010) argue that the ability construct underlying the assessment must be linked to this language use (e. g. assessment must be based on a language needs analysis). The tasks used in the assessment must have relevant properties in common with the tasks found in the domain of target-language use, and scoring performance must reflect the aspects that are emphasised in the construct definition. In addition, the construct and requirements must be clearly communicated to all relevant stakeholders, including the test takers. The claim to meaningfulness touches upon aspects that were formerly often discussed under the heading of *construct validity*.

The test construct is also an issue in the literature covered by this paper. In their discussion of the use of language tests for immigration and integration purposes, Saville & Van Avermaet (2008) present key issues that testers must address when developing suitable language tests; one of them is the test construct. Saville & Van Avermaet argue that, because the test is officially framed as instrumental in promoting the applicant's integration, the construct must be defined accordingly. That means the test must provide information on the immigrant's ability to function in the host society (rather than merely assessing decontextualised linguistic knowledge). The language proficiency level targeted should also conform to the designated purpose of the test. In other words, if the purpose of the test is to assess how socially integrated a test taker is, then the expected proficiency level should correspond to the real-world demands posed by actual language used in everyday social life. In concrete terms, the guidelines developed by the ALTE LAMI group (Balch et al., 2008) postulate that the real-world demands should be identified by means of a needs analysis which investigates the practical situations and activities the test takers (here: prospective residents or citizens) will be facing in their lives. The identification of real-world demands then enables test developers to determine, in a next step, what language skills are to be assessed and to define the desired proficiency level in the target language. The LAMI group suggests that the communication requirements should not be established for 'immigrants' in general, but be defined specifically for different subgroups (e. g. working immigrants, spouses, etc.) as their respective real-world needs may be different. Shohamy (2009) points out that tests themselves often focus on unrealistic language standards (e. g. linguistic correctness) that fail to incorporate the specificities of second language use and daily multilingual practices. Language tests therefore do not reflect actual real-world needs that immigrants experience in the host country, which places their suitability as a meaningful point of reference for decisions in question.

Fulcher (2004) observes that many countries that introduced a language requirement as part of the immigration and integration procedure defined the required level of proficiency in relation to the CEFR scale (North, 2009; Council of Europe, 2001). This practice is a rather controversial issue in the literature analysed here.

The controversy concerning the use of the CEFR concerns mainly, but not exclusively,14 the construct definition. Krumm (2007) draws attention to the fact that the CEFR reference levels were not initially designed as a basis for assessing the language skills of immigrants, but rather those of classic foreign-language learners. Krumm (2007) provides the example of a descriptor at level A1 that contains the element 'propose a toast', which is obviously not a very relevant skill for the majority of immigrants for whom the language requirements

14 | Extra et al. (2009b) and also Yoffe (2010), for example, criticize the use of the CEFR level descriptors as a prescriptive tool for immigration and integration purposes because, according to them, the illustrative scales of the CEFR were originally developed as a descriptive tool to be used to acknowledge people's language skills. And, while the CEFR was created to encourage international mobility, its instrumentalisation in immigration policies has rather a contrary effect: it makes the CEFR a part of a system of social exclusion and gate-keeping.

were put in place. In line with the suggestions of the LAMI group, Krumm suggests that the test construct and the corresponding descriptors be adapted to those contexts of language use that immigrants actually encounter. Also, the fact should be acknowledged that, in their multilingual reality, immigrants often know and use a range of other languages in addition to their mother tongue, and adequate descriptors should integrate intercultural and plurilingual dimensions. Krumm indicates another facet of plurilingualism that should be accounted for, as it frequently arises among immigrants: an uneven proficiency profile within one and the same language. Immigrants often develop differing skills in a language, e. g. low writing but good reading and listening skills. Krumm (2007) concludes that the CEFR should be adapted to the specific context of immigration by doing more development work on partial competence and plurilingual repertoires. In view of assessing such language skills, he suggests that the European Language Portfolio might be better suited to accommodate variable linguistic repertoires than a test (see also Papp, 2010).

Shohamy (2007) is similarly critical regarding the suitability of the CEFR as a basis for construct definition in an immigration context. One point of criticism concerns the way the higher CEFR proficiency levels are described. These descriptions make assumptions about knowledge, skills and communicative functioning that cannot be generalised in reference to the immigrant population. Shohamy also criticises that the descriptors inadequately reflect the context for this target group – the purpose of the assessment, the age of the learners, the varying learning contexts, or the functional distribution of the different languages individuals know.

Alderson (2007) calls attention to a slightly different problem that arises when using the CEFR. He observes a tendency in politicians and civilians to define the CEFR proficiency standards to be met by immigrants without consulting language experts for their recommendations. Cooke (2009) provides the example of the UK citizenship test, for which the level of proficiency ('Entry 3' on the Adult ESOL core curriculum progression, equivalent to CEFR level B1) was decided upon by the government. Following this decision, the set level was heavily criticised by ESOL practitioners as too high. The situation is similar in Germany where the government also set the language requirement for permanent residency and citizenship at proficiency level B1 (Plassmann, 2011) – a decision that, again, has been criticised for being unrealistic (Laversuch, 2008).

Some studies fundamentally call the language proficiency constructs used into question. Shohamy's radical view that language tests for immigration touch upon a person's right to use their own language has been mentioned above. From a more practical stance, Shohamy (2009) questions the assumption that proficiency in the national language is actually needed to function in society. And in case it is needed, the extent and level of proficiency required should be reconsidered. In a similar vein as Shohamy, Saville & Van Avermaet (2008) put the importance of knowing the local language into perspective by referring to the fact that the difficulty immigrants face in accessing employment and education does not necessarily stem from a lack language skills, but is rather a consequence of the marginalisation they might suffer as foreigners from the outset. They also call to mind that these people, as they live in multilingual societies, might function in their everyday lives using languages other than the locally official language, but speaking these languages is not recognised as participating in social life. Slade (2010a) points out that the linguistic and cultural knowledge immigrants acquire over time and through their individual experiences can vary greatly, making it difficult to establish (well-founded) standards to which all immigrants should aspire.

Papp (2010) reports on a small validity study she performed on the British test for citizenship and residency 'Life in the UK' (introduced in 2005). She examined the test itself and the materials available for prospective candidates. The study focuses in particular on whether or not the test materials reflect the domain and proficiency level of target language use ('functional competence required for the successful demonstration of citizenship and settlement'). Papp's work yielded two main findings. First, the analysis of the test and documentation shows that the language ability tested does not provide adequate evidence of the test taker's ability to integrate into the host society. The contents of the test furthermore do not correspond to the domain of language use relevant for participating in everyday social life (e. g. employment, public administration, education). Second, by applying corpus analysis, Papp arrived at the conclusion that the level of language proficiency of both test and support materials is higher than the targeted level. This rather negative overall result highlights a problem of meaningfulness (and construct validity) as the test design and the interpretation of the assessment records do not accurately reflect the initial test construct.

The report on the Australian Citizenship Test (ACT) was compiled by a committee chaired by Woolcott, a diplomat, on account of the frequent criticism of the test. The report focused on the appropriateness of the language level required by the test in addition to disputed content and test bias. According to the citizenship law, the test should be designed to assess *basic* knowledge of the English language. In the submissions language experts made on behalf of the Woolcott committee, it was criticised that the language demands as embodied by the English language used in the preparation booklet and the formulation of the test items went well beyond 'basic knowledge'. The Woolcott Report attempted to define the requirement of 'basic knowledge' more meaningfully as 'having sufficient knowledge of English to exist in the wider Australian community' and locating it in the vicinity of the CEFR level band of A1/A2 (McNamara & Ryan, 2011). Apparently this clarification was of little practical consequence: McNamara & Ryan (2011) state that the level of English used in the revised test and booklet remained virtually unchanged.

Kunnan (2009) critically compares the declared purpose (serving as an implicit construct) of the redesigned US Naturalization Test, instated in 2008, to its actual implementation. The designated purpose of the test is to promote civic participation and social integration. Kunnan took issue with two main points. One concerns the type of civic knowledge addressed by a test that only refers to memorised facts instead of requiring tasks that encourage and evaluate actual understanding of US history and the governmental system. The second point of concern is the type and level of knowledge of English required to pass the test. According to Kunnan (2009), the level is probably too low to sufficiently demonstrate that a person is able to participate in civic and social life in the US. In light of these and several other weaknesses, Kunnan (2009) comes to the conclusion that the U.S. Naturalization Test cannot serve its declared purpose regarding civic and social integration and can therefore not be rated as a meaningful test:

> The Naturalization Test as it is conceptualized cannot test civic nationalism or social integration through indirect measures of English language ability and knowledge of U.S. history and government, as these are skills and knowledge but not measures of community participation and activism (Kunnan, 2009: 95).

Kunnan names an additional, negative factor in the redesigned Naturalization Test: the low naturalisation rates, which indicate that the test may actually be discouraging potential candidates.

Only few articles are available on the official German Test for Immigrants (DTZ) written by persons who were actively involved in the test's development. Perlmann-Balme (2011) and Plassmann (2011), both, seek to demonstrate that their work corresponds to the quality standards defined by ALTE, the association to which most important test providers in Europe belong. ALTE Standards 1 and 2 are related to the meaningfulness of test-based interpretations as pertaining to target-language use:

> 1) The examination is based on a theoretical construct, e. g. on a model of communicative competence.
> 2) You can describe the purpose and context of use of the examination, and the population for which the examination is appropriate (ALTE [Association of Language Testers in Europe], 2007).

The developers of the DTZ attempt to meet these standards by combining a generally action-oriented approach as described by the CEFR (Council of Europe, 2001) that focuses separately on language proficiency in the four classic language skills, with an orientation to the specific needs of the target group(s). The German framework curriculum for integration courses (Buhlmann et al., 2007) serves as the source for information on specific needs. For the DTZ construct, the core domains of language use common to the different test taker groups are extracted from the 12 domains outlined in the source document. The construct and the goals of the DTZ are communicated to the general public by means of the DTZ handbook (Perlmann-Balme et al., 2009), which also contains a sample test.

In their articles, Perlmann-Balme and Plassmann explicate the steps taken to produce a test that puts a meaningful construct into practice while adhering to the ALTE standards. Unfortunately, they do not provide enough evidence for a reader to come to an independent conclusion.

De Jong et al. (2009) make a much greater effort to demonstrate that their telephone and computer-based Test of Spoken Dutch (TGN) allows for meaningful interpretations in the context of testing for immigration and integration purposes, even though it uses mostly item types that would not intuitively be identified with communicative speaking tasks occurring in real life. Three item types are used for the TGN: sentence repetition, short-answer questions (which come closest to an authentic task type), and opposites of single words15. There is no human interlocutor; the entire test is administered and scored by a computer. All item types include spoken input and some spoken output. According to the authors,

> the test measures the facility with which candidates are able to track what is said, extract meaning in real time, and formulate and produce relevant, intelligible responses, at a conversational pace (De Jong et al., 2009: 43).

15 | At the end of each test, two story retelling items are given. These, however, serve test validation and research purposes only.

Presenting the item demands in this manner evokes parallels with cognitive and linguistic operations that also play an important role in spoken interaction.
The authors present more arguments for the sake of demonstrating the proximity between language use in the test and actual target language use:

→ The stimuli used for sentence repetition tasks were selected from a corpus of spoken Dutch and represent everyday spontaneous speech from different regions.
→ The automatic scoring system was compiled and trained using a large sample of native and non-native speakers of Dutch in order to represent the range of Dutch that is used in real life.
→ The TGN assesses vocabulary, grammar, pronunciation and fluency as indicators of spoken language proficiency. Actual performance assessments also commonly rate the same features of a candidate's language, as these skills are considered fundamental.

A big difference that remains between the two types of testing, but also between language use on the TGN and language use in real-life situations, is the degree of openness in the tasks. While the three TGN item types narrowly guide language (re-)production, language use in target situations often calls for active construction and negotiation of discourse.

Obviously, the developers of the TGN anticipated this potential weakness in their validity argument and therefore had a large number of candidates carry out additional tasks such as story retelling, answering open questions and participating in an oral interview. The results were assessed by humans, some of them in relation to the CEFR levels of language proficiency, so that all the results could be incorporated in a common analysis culminating in a single CEFR-related scale. Subsequently, this scale could be used to relate the TGN scale that is based on machine ratings of vocabulary, grammar, pronunciation and fluency with satisfactory precision to the CEFR, especially in the lower scale regions where the important pass-fail decisions are taken.

Not surprisingly, the topic of meaningfulness of test-based interpretations – which covers much of what was formerly treated under the heading of construct validity – receives considerable attention in the literature covered in our study. Some authors deal with the question of what an appropriate test construct should be, whether the construct should extend beyond a single language, whether the CEFR proficiency scales make sense in this concrete context, or whether any requirements can be justified at all. Others critically investigate existing tests, and still others explicate and justify the constructs they use for their own test. In order to illustrate the latter case, the argument regarding the construct of the Test of Spoken Dutch was outlined in more detail than others in our study because it is quite exceptional in its thoroughness and transparency. It should not be forgotten, however, that, despite its qualities, this paper is overall an affirmative text written by test developers wishing to demonstrate that their test suits its purpose.

3.7 Impartiality of interpretations for all groups of test takers

According to Bachman & Palmer (2010), interpretations of assessment data are only impartial if all aspects of test administration, including the formats and contents of an assessment as well as access to information on the assessment, do not specifically disadvantage any group of candidates.

De Jong et al. (2009) as well as Perlmann-Balme (2011) and Plassmann (2011) – those publication in our collection that actually provide some detail on test development – address aspects of impartiality.

De Jong et al. (2009) apparently consider test administration in Dutch embassies around the world a potential problem area that might introduce bias. In their article they defend or justify this practice using the following arguments. 1) The explanations on testing procedures are sufficient for every test taker because they are generally communicated by trained personnel in the language of the test taker. Where no such person is available, the test takers are free to bring an interpreter to the embassy. 2) While fraud is a well-known source of partiality in highly decentralised test systems, the developers claim it is virtually impossible on the TGN because the actual test is centrally provided on the basis of a large item bank, and because it is unique for every test taker and automatically scored by a computerised system.

The TGN developers also make an effort to demonstrate that item formats and content do not pose a threat to the impartiality of their test. One of the item formats used on the TGN is short answers. These are necessarily open-ended to some degree and involve previous content knowledge. This may result in test bias when the test is administered internationally to a very mixed audience. The TGN developers counter this suspicion by demonstrating that they minimised this possibility by pre-testing their items on three different groups: potential test candidates from immigrant schools; immigrants from outside the schools; and native speakers from different social backgrounds and age groups, men and women alike. Similarly, the developers of the TGN tested several other hypotheses relating to potential background variables that might unduly influence success on the TGN, e. g. achieved level of education, age, gender, degree of literacy, and also possible over-punishment of strong accents through the automatic scoring system.

In relation to ALTE standard 2 – 'You can describe … the population for which the examination is appropriate' - the developers of the German DTZ mention measures taken to minimise potential unequal treatment of members of the target groups taking their test. These groups include, on the one hand, immigrants with a modest educational background who want to be active in the family or find a job requiring minimal qualifications, and, on the other hand, immigrants with significant learning experience, even diplomas, who wish to pursue a career. One of the measures is related to test content and construct: tasks were selected only from those 'core' domains described in the framework curriculum thought to be relevant for all groups. The other main measure taken concerns the types of reviewers and experts who were involved during test development, specifically teachers of the target group, teacher trainers and textbook authors. It was the duty of these professionals to ensure that none of the groups is at a disadvantage due to inappropriate topics and content, but also on account of item formats or elements of test administration. The information provided by these experts was additionally matched with results from the statistical analyses of trial runs of the DTZ.

3.8 Consistency of assessment records

The degree of consistency in the information gained from an assessment determines how reliably interpretations can be made on the basis of such information. Ideally, the data contain no systematic variation due to differences in test administration, test scoring or test-taker group. The issues formerly treated under the heading of *reliability* fall under these points. Therefore, it is essential for test developers to demonstrate that their assessment records are consistent. Formal examinations are normally more reliable and objective than informal assessments and, as mentioned in chapter 3.5, even authors sceptical of testing appreciate the increased fairness gained by using them (Balch et al., 2008).

Considering the long-standing tradition reliability analyses have, it is hardly surprising that consistency-related points are presented at some length in De Jong et al. (2009), Perlmann-Balme (2011) and Plassmann (2011).

De Jong et al. (2009) introduce their automated system of test administration and scoring, including the test item bank, as one important factor that helps to generate consistent test data; its advantages stem from its ability to reduce the possibility for humans to introduce disturbances or bias. But, the great challenge for the developers of the TGN consists in demonstrating that users actually have good reasons to believe that the machine-based system works reliably. A considerable portion of the article is dedicated to achieve just this. De Jong et al. (2009) provide evidence of high correlations between human and machine scoring. They also show that the measurement of the ability in question is possible with only a small statistical error in the relevant lower section of the ability scale where the pass-fail cut-off is located. The developers' argument is strengthened by impressive numbers of test persons and items that generated the data for the analyses.

The German DTZ developers for their part use ALTE standards as a structuring element in the presentation of their work on the 'reliability argument' (Perlmann-Balme, 2011).

ALTE standard number 3 is related to good practice in test construction:

> 3) You provide criteria for selection and training of test constructors[,] and expert judgement is involved both in test construction, and in the review and revision of the examinations (ALTE [Association of Language Testers in Europe], 2007).

In their report on the DTZ development, the German authors elaborate on writer and reviewer training, and on their interactive and stepwise development methods that combined trial runs (pilot testing), teacher feedback and statistical analysis until the actual test papers could be produced with confidence.

ALTE standard number 4 calls for the equivalence of exams administered on different occasions:

> 4) Parallel examinations are comparable across different administrations in terms of content, stability, consistency and grade boundaries (ALTE [Association of Language Testers in Europe], 2007).

Perlmann-Balme (2011) refers to the tagged DTZ item bank as the technical basis that helps to guarantee comparability of exam content across different versions, while exam regulations and implementing rules assure organisational consistency across test administrations. The examination handbook (Perlmann-Balme et al., 2009) contains a transcription of an oral exam and plays an key role in standardising the course and the assessment of that exam. Perlmann-Balme (2011) and Plassmann (2011) both emphasise the high quality of the procedures developed to assess the written texts: all written texts are rated centrally, which, according to Plassmann, has the great advantage that problems can be treated by the right persons as they arise. In order to assure consistent ratings, three concrete measures are in place: rater training; double ratings; post-analyses of ratings in order to detect outlier judgments.

ALTE standard number 14 requires statistical item analysis:

> 14) Item-level data (e. g. for computing the difficulty, discrimination, reliability and standard errors of measurement of the examination) is collected from an adequate sample of candidates and analysed (ALTE [Association of Language Testers in Europe], 2007).

Perlmann-Balme (2011) just writes briefly that the standard psychometric analyses mentioned in standard 14 were actually carried out. Also, items were eliminated on the basis of the results.

This brief chapter on the consistency of assessment records once again demonstrates, as did the previous chapters, that the publications that emerged from high-stakes projects in the Netherlands and Germany actually deal with central elements of standard test validation procedures. The more research-oriented article by De Jong et al. (2009) proceeds quite differently from the approach taken by the authors writing about the German DTZ. While De Jong et al. (2009) makes a clear effort to demonstrate measures taken to ensure sufficient validity, Perlmann-Balme (2011) and Plassmann (2011) tend to simply declare that they adopted high standards and followed the right procedures. This difference may be entirely due to the style and target audience of the journals that published these articles. From the point of view of an expert reader, a more thorough follow-up to the articles on the DTZ would be highly desirable.

Chapter 3 overall investigates how the literature covered by this study deals with questions of assessment validity and validation. The aspects of validity we looked into more closely form part of Bachman & Palmer's assessment validation framework (AUA) (Bachman & Palmer, 2010). Our analyses reveal that most authors focus on only one or very few aspects of validity and validation. Much interest seems to lie in fundamental issues as well as in socio-political considerations related to testing for immigration and integration purposes. These topics are only partly covered even by up-to-date validation frameworks with an orientation towards test use and test impact, like the AUA. Discussions of ideologies implicit in test constructs (e. g. 'one nation – one language') or considerations related to the 'justice' or ethics of testing immigrant groups lie beyond the scope of conceptual frameworks for assessment validation – and also beyond the area of expertise and experience of language testers.

One validity criterion, 'meaningfulness of interpretations with respect to a construct', otherwise often treated under the heading of *construct validity*, has gained the attention of both test developers and authors sceptical of the tests. While test critics challenge the link between test content and (real-world) target-language use, test developers make an effort to demonstrate or explicate the meaningfulness of their tests.

We were surprised to see how few publications in the field of language testing for immigration can actually be considered validation studies. True validation studies are either generally scarce, or they are not made available to an audience of independent assessment experts.

4

Summary and conclusions

This study provides an overview of the recent developments in the field of language testing for immigration and integration purposes, the test validity-related investigations and discussions going on in this context, and, implicitly, the issues at stake and the arguments used. The information gained forms the basis for the following concluding observations and proposals.

Our findings reveal several rapid and quite drastic developments in the field of language testing for immigration and integration purposes in Europe and other, mainly Western, countries over the past decade. The general tendencies observed are that language requirements are increasingly imposed in more steps in the immigration and integration process – in some countries even before actual immigration occurs – and that the requirements have become more comprehensive, with a discernable trend towards formal assessment, often standardised tests. In some countries the assessments not only cover language but also cultural knowledge. Typically for Europe, the new requirements do not apply equally to all groups of immigrants, due to EU-related regulations.

Apart from common general trends, we were able to identify considerable diversity in the concrete design and definition of the requirements and the corresponding assessments. Variable elements include: status of preparatory language and integration courses; exemption of certain types of immigrants; test takers' minimal age; fees; test design and administration; test methods; skills tested; and required level of language proficiency. As an example, proficiency levels required for admission to a country range between A1 and B1 on the European reference scale; for residency and citizenship the levels vary between A1 and B2. The rationale given for the choice of levels, however, is usually very similar.

The comprehensive and widely known Assessment Use Argument (AUA) framework by Bachman and Palmer (Bachman & Palmer, 2010) was used to summarise the discussion of validity-related topics in the literature on language testing for immigration. This validation framework for language assessments is comprehensive in that it takes into account and conceptually links all aspects of an operational test that may be relevant to its validity – from the test tasks and administration to the impact of a test on the individuals concerned and on society at large.

The publications included in the present synthesis study provide a rather uneven treatment of the various topics related to the valid use of tests. Only very few papers are actually intended to comprehensively demonstrate (or systematically challenge) the validity of a specific test for immigration. Those that actually do focus on a specific test system were written by the developers in order to justify test use. One of these papers (De Jong et al., 2009) proceeds very systematically in its attempt to provide enough validity evidence to justify the actual use of the Test of Spoken Dutch (TGN). It is obvious that the authors follow an argument-based validation framework like the one presented in Bachman & Palmer (2010) although this is never explicitly stated. Most papers in our selection deal with either the *meaningfulness* of test content and levels, the *beneficence* a test brings to the persons concerned or even raise fundamental questions about the testing of immigrants *per se* that can hardly, if at all, be accommodated by existing validation frameworks.

Several authors doubt the *meaningfulness* of interpretations based on language (and/or cultural knowledge) test results because they have diagnosed or suspect a mismatch between

- → test content and real-world needs and tasks;
- → the declared test construct (e. g. 'sufficient language knowledge for social integration') and the skills and competences actually tested (more or less arbitrarily determined);
- → the level of competence required for the test and for real-life activities;
- → the language tested and the language(s) needed and used by the immigrants in the multilingual social contexts in which they live and work;
- → or, even more fundamentally, success on lan-

guage or/and cultural knowledge tests and good citizenship.

When (stricter) testing regimes are imposed, the authorities in charge often claim that the effects of testing are mostly *beneficial*, not least for the immigrants themselves. This point is touched upon in a number of studies, quite often merely in an anecdotic manner, but in some cases also based on actual impact studies like the INTEC study (Strik et al., 2010). Negative appraisals of the test benefits predominate in the publications examined. The obligation to pass (stricter) tests in order to move forward in the immigration process is often judged as useless, stressful for the immigrants and discouraging, or it is criticised for selectively excluding certain groups (e. g. individuals with little formal education) from fuller integration. Some appreciation is voiced for the preparatory courses that sometimes accompany language and cultural knowledge tests; the main benefits identified are that such courses provide an (compulsory) opportunity for immigrants to leave the confines of the own community and enter into contact with other people. The INTEC study based on statistical data and interviews from nine European countries corroborates other, more or less informal findings of a dissuasive effect of new requirements and very few beneficial effects on individuals.

Critical language testers such as Shohamy and McNamara contribute a more radical perspective to the discussion on language testing for immigration and integration purposes by asking *fundamental questions* (questions of 'justice' according to McNamara and colleagues) that touch upon several fields other than language testing including (critical) social and political science, philosophy (ethics) and constitutional law. By uncovering implicit ideologies ('one nation – one language'), 'hidden agendas' (language tests for gate-keeping purposes), political motives (securing social cohesion through exclusion), ethically problematic consequences of tests (exclusion from participation in society), these critical language testers instigate a far-reaching debate involving a number of disciplines that have thus far hardly been drawn on for the development and implementation of language and cultural knowledge tests. Indeed, experts from these fields would have to be involved long before a test was actually commissioned – if a test must be introduced at all.

What conclusions can be drawn from our observations? What directions for the future can be derived?

It seems desirable that the issues raised by the authors critical of language testing become part of the discussions guiding actual language policy and assessment practices, as their objections are undoubtedly fundamental in nature. We believe, however, that no changes will be initiated without a considerable effort by academics from disciplines that do not usually deal with language requirements and testing. Decisions on language requirements and tests for immigration and integration purposes are often taken by legislative powers, and implementation is commissioned and supervised by state authorities. These bodies are not necessarily obligated to invite critical scholars to discuss and settle 'issues of social value', as McNamara & Ryan (2011) suggest. States are naturally only interested in these issues when enforceable rights are (potentially) violated. In lieu of governmental action, the academic community is called on to take the steps necessary to autonomously enlarge the scope of the discussion and to bring the crucial topics to the fore. In addition, to have sufficient impact, it appears necessary to involve the disciplines that normally deal with the social, legal and ethical issues at stake.

One point that emerges quite clearly is the potential usefulness of a validation framework such as Bachman & Palmers Assessment Use Argument (AUA). From a professional standpoint, language testers can do no less than accept accountability as a guiding principle or 'axiom' (cf. Bachman & Palmer, 2010). A validation framework operationalises accountability to ensure state-of-the-art test development and responsible test use. Authorities that (potentially) commission tests should be made increasingly aware of the existence and the scope of up-to-date frameworks such as the AUA; this enables the persons responsible to formulate calls for tender and project evaluation criteria accordingly and to be prepared to allocate the necessary time and finances. Moreover, comprehensive validation frameworks that cover everything from a candidate's preparation for a test to the test's impact on society make clear – especially to those who put testing systems in place and are responsible for them – that test validation is an ongoing process that extends over the whole life span of a test. The idea of a one-time development phase is not sustainable.

As mentioned above, the meaningfulness of interpretations based on scores from the language tests is a frequently contested point in the publications surveyed. In particular, several authors put forth the claim that the abilities assessed and the abilities necessary to function in and contribute to society have little to do with each other. The main measure to achieve better agreement – and also to overcome expectations regarding language knowledge that are based on ideologies – is a thorough empirical needs analysis[16]. Needs analysis was one of the methods applied successfully in the Swiss *fide* project (Müller & Wertenschlag, 2013). It included not only immigrants from different social and ethnic groups but also members of the host society who have regular contact to immigrants with limited language proficiency. A further critical matter is ensuring an adequate interpretation of the results of needs analyses. They may reveal, for example, that the need to know a local language (at a specific level) differs very much according to group membership; that it makes little sense to require certain individuals to acquire written language skills; or that some immigrants would actually need to have very good command of a local language in order to be able to contribute to society according to their abilities. An assessment and certification system that takes differentiated insights into account could opt for a modular approach to language assessment, thus responding to differential needs and abilities. With integration in mind, it might be a good idea to start with designing an optimal language support system and to move towards certification from there. The *fide* system supported by the Swiss authorities (Gerber & Schleiss, 2013) seeks to react to diversity by accommodating differing candidate needs and profiles, and by embedding language assessment in a system of language support[17].

Another insight emerging from our study is a need for more impact studies that cover a wider range of issues in more contexts. In many of the articles, uncertainties about the actual impact of language requirements and language tests become apparent. At times, speculation and anecdotal evidence supplant well-founded evidence. The INTEC Study (Strik et al., 2010) has already achieved a great deal by uniting quantitative and qualitative data on the impact of testing regimes from nine European countries; smaller-scale studies provide additional knowledge about other contexts. These efforts need to continue and to be extended to other contexts and research questions. According to current validation frameworks like the AUA – that understand test use as a social phenomenon and that expect tests to benefit stakeholders and society overall – impact studies form a constitutive element of comprehensive test validation that contributes considerably to a well-founded justification of its use.

In our final recommendation, we would like to stress that it is highly desirable that tests for immigration and integration purposes be validated not only more systematically but also more transparently. In our collection of published literature, only one article (De Jong et al., 2009) is actually geared towards delivering a comprehensive validation argument by addressing many relevant points and by providing arguments and evidence. It must, however, be noted that the orientation of this article is confirmatory[18]. As might be expected, it was published at a stage when the test was already operational because, as Briggs notes

16 | Language needs analysis may be extended to a series of language audits that, within relevant contexts, take into account not only the various stakeholders' needs but also their multilingual language abilities.

17 | Similarly, Schneider et al. (2006) recommended defining two language proficiency profiles in the context of naturalisation in Switzerland: a support profile and an assessment profile.

18 | Chapelle et al. (2008) is another very worthwhile report written by (some of) its developers from a different context, namely, the TOEFL iBT, a relatively new test for prospective students at English-speaking universities. The authors use a validation framework that borrows from Kane and Bachman & Palmer.

in his commentary on Kane's (Kane, 2004) call for an *explicit* interpretive argument (similar to Bachman and Palmer's AUA), there may be a 'paradoxical' real-world problem involved that cannot be easily overcome: if a test has not been fully validated, then using it to make high-stakes decisions becomes questionable. But, if it has not been administered to a sample from the target population for the expressed purpose of the test, it becomes impossible to fully validate the test (Briggs, 2004). In our opinion this 'paradox' does not exempt test developers and those who use and/or are responsible for a test by any means from conducting validation studies and making them available to an independent expert audience; at most, it provides justification for slightly deferred publication. A desirable next step, now that good quality and practically manageable validation frameworks are available, is to create a 'validation argument culture' in which publishing validity investigations on operational tests is the norm and not the exception. Also, in order to counteract the bias introduced by the developers-as-authors, external testing experts should be commissioned to investigate sensitive points in a test system (including impact) and given access to the inner workings and confidential data if needed. Dedication to test validity, a high degree of transparency – and of course favourable results from validation studies – would considerably increase the credibility and legitimacy of an institution – even state authorities – to assess language competencies in such a sensitive area as immigration.

5

Bibliography

A

Adami, H. & Leclercq, V. (Eds.). (2012). *Les migrants face aux langues des pays d'accueil*. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Adamo, S. (2008). Northern exposure: the new Danish model of citizenship test. *International Journal on Multicultural Societies*, *10*(1), 10–28.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*(4), 659–663. doi:10.2307/4626093

ALTE [Association of Language Testers in Europe] (Ed.). (1994). The ALTE code of practice. Retrieved from http://www.alte.org/attachments/files/code_practice_eng.pdf

ALTE [Association of Language Testers in Europe]. (2001). Principles of good practice for ALTE examinations. Revised draft. Retrieved from http://www.alte.org/attachments/files/good_practice.pdf

ALTE [Association of Language Testers in Europe] (Ed.). (2007). Minimum standards for establishing quality profiles in ALTE examinations. Retrieved from http://www.alte.org/attachments/files/minimum_standards.pdf

American Educational Research Association, American Psychological Association, National Council on Measurement in Education & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

B

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

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1–34.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford [etc.]: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice: developing language assessments and justifying their use in the real world*. Oxford [etc.]: Oxford University Press.

Balch, A., Corrigan, M., Gysen, S., Kuijper, H., Perlmann-Balme, M., Roppe, S., … Zeidler, B. (2008). *Language tests for social cohesion and citizenship - an outline for policymakers*. Strasbourg: Council of Europe.

Blackledge, A. (2009a). "As a country we do expect": the further extension of language testing regimes in the United Kingdom. *Language Assessment Quarterly*, *6*(1), 6–16. doi:10.1080/15434300802606465

Blackledge, A. (2009b). Being English, speaking English: extension to English language testing legislation and the future of multicultural Britain. In G. Hogan-Brun, C. Mar-Molinero & P. Stevenson (Eds.), *Discourses on language and integration: critical perspectives on language testing regimes in Europe* (pp. 83–108). Amsterdam and Philadelphia: John Benjamins Publishing Company.

Blackledge, A. (2009c). Inventing English as convenient fiction: language testing regimes in the United Kingdom. In G. Extra, M. Spotti, & P. Van Avermaet (Eds.), *Language testing, migration and citizenship: Cross-national perspectives on integration regimes* (pp. 66–86). London; New York: Continuum International Publishing Group.

Böcker, A. & Strik, T. (2011). Language and knowledge tests for permanent residence rights: help or hindrance for integration? *European Journal of Migration and Law*, *13*(2), 157–184. doi:10.1163/157181611X571268

Brennan, R. L., National Council on Measurement in Education & American Council on Education (Eds.). (2006). *Educational measurement* (4th edition.). Phoenix, AZ: Greenwood.

Briggs, D. C. (2004). Comment: making an argument for design validity before interpretive validity. *Measurement: Interdisciplinary Research & Perspective*, *2*(3), 171–174. doi:10.1207/s15366359mea0203_2

Buhlmann, R., Ende, K., Kaufmann, S., Kilimann A. & Schmitz, H. (2007). *Rahmencurriculum für Integrationskurse Deutsch als Zweitsprache*. München: Goethe-Institut. Retrieved from http://www.bamf.de/SharedDocs/Anlagen/DE/Downloads/Infothek/Integrationskurse/Kurstraeger/KonzepteLeitfaeden/rahmencurriculum-integrationskurs.pdf

C

Chapelle, C. A., Enright, M. K. & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York and London: Routledge.

Cooke, M. (2009). Barrier or entitlement? The language and citizenship agenda in the United Kingdom. *Language Assessment Quarterly*, *6*(1), 71–77. doi:10.1080/15434300802606580

Council of Europe (Ed.). (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Cox, L. (2010). The value of values? Debating identity, citizenship and multiculturalism in contemporary Australia. In C. Slade & M. Möllering (Eds.), *From migrant to citizen: testing language, testing culture* (pp. 77–97). Basingstoke: Palgrave Macmillan.

D

De Jong, J. H. A. L., Lennig, M., Kerkhoff, A. & Poelmans, P. (2009). Development of a test of spoken Dutch for prospective immigrants. *Language Assessment Quarterly*, *6*(1), 41–60. doi:10.1080/15434300802606564

E

Educational Testing Service (ETS). (2002). ETS standards for quality and fairness. Educational Testing Service. Retrieved from http://www.ets.org/s/about/pdf/standards.pdf

Extra, G. & Spotti, M. (2009a). Language, migration and citizenship: a case study on testing regimes in the Netherlands. In G. Hogan-Brun, C. Mar-Molinero & P. Stevenson (Eds.), *Discourses on language and integration: critical perspectives on language testing regimes in Europe* (pp. 61–81). Amsterdam and Philadelphia: John Benjamins Publishing.

Extra, G. & Spotti, M. (2009b). Testing regimes for newcomers to the Netherlands. In G. Extra, M. Spotti & P. Van Avermaet (Eds.), *Language testing, migration and citizenship: cross-national perspectives on integration regimes* (pp. 125–147). London and New York: Continuum International Publishing Group.

Extra, G., Spotti, M. A. & Van Avermaet, P. (Eds.). (2009a). *Language testing, migration and citizenship: cross-national perspectives on integration regimes*. London and New York: Continuum International Publishing Group.

Extra, G., Spotti, M. & Van Avermaet, P. (2009b). Testing regimes for newcomers. In G. Extra, M. Spotti & P. Van Avermaet (Eds.), *Language testing, migration and citizenship: Cross-national perspectives on integration regimes* (pp. 3–33). London and New York: Continuum International Publishing Group.

Extramania, C. (2012). Les politiques linguistiques concernant les adultes migrants: une perspective européenne. In H. Adami & V. Leclercq (Eds.), *Les migrants face aux langues des pays d'accueil: acquisition en milieu naturel et formation* (pp. 135–152). Villeneuve d'Ascq: Presses Univ. Septentrion.

Extramania, C. & Van Avermaet, P. (2010). *Language requirements for adult migrants in Council of Europe member states: report on a survey*. Strasbourg: Council of Europe. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Mig-ReportSurvey2011_EN.doc

F

Farrell, E. (2010). "Do I feel Australian? No you tell me": Debating the introduction of the Australian formal citizenship test. In C. Slade & M. Möllering (Eds.), *From migrant to citizen: testing language, testing culture* (pp. 164–187). Basingstoke: Palgrave Macmillan.

Fulcher, G. (2004). Deluded by Artifices? The Common European Framework and Harmonization. *Language Assessment Quarterly*, *1*(4), 253–266. doi:10.1207/s15434311laq0104_4

G

Gerber, A. & Schleiss, M. (2013). Langue et intégration: une responsabilité partagée / Sprache und Integration: eine gemeinsame Verantwortung. *Babylonia*, (1), 9–12.

Goodman, S. W. (2011). Controlling immigration through language and country knowledge requirements. *West European Politics*, 34(2), 235-255. doi:10.1080/01402382.2011.546569

H

Hargreaves, M. (2010). Citizenship testing in the anglophone countries: The UK, Canada and the USA. In C. Slade & M. Möllering (Eds.), *From migrant to citizen: testing language, testing culture* (pp. 101–124). Basingstoke: Palgrave Macmillan.

Hogan-Brun, G., Mar-Molinero, C. & Stevenson, P. (Eds.). (2009a). *Discourses on language and integration: critical perspectives on language testing regimes in Europe*. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Hogan-Brun, G., Mar-Molinero, C. & Stevenson, P. (2009b). Testing regimes: Introducing cross-national perspectives on language, migration and citizenship. In G. Hogan-Brun, C. Mar-Molinero & P. Stevenson (Eds.), *Discourses on language and integration*. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Holland, A. (2010). Australian citizenship in the twenty-first century: Historical perspectives. In C. Slade & M. Möllering (Eds.), *From migrant to citizen: testing language, testing culture* (pp. 39–59). Basingstoke: Palgrave Macmillan.

I

ILTA (International Language Testing Association). (2007). Guidelines for practice. Retrieved from http://www.iltaonline.com/images/pdfs/ILTA_Guidelines.pdf

J

Joppke, C. (2010). How liberal are citizenship tests? *EUI Working Papers*, *41*, 1–4.

K

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective*, *2*(3), 135–170. doi:10.1207/s15366359mea0203_1

Kane, M. (2006). Validation. In R. L. Brennan, National Council on Measurement in Education, & American Council on Education (Eds.), *Educational measurement* (4th edition., pp. 17–64). Phoenix, AZ: Greenwood.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17. doi:10.1111/j.1745-3992.1999.tb00010.x

Kiwan, D. (2008). A journey to citizenship in the United Kingdom. *International Journal on Multicultural Societies*, *10*(1), 60–74.

Klein, G. (2013). Do gender, age and first language predict the results in the Deutsch-Test für Zuwanderer (DTZ)? In E. Dimitrova-Galaczi & C. J. Weir (Eds.), *Exploring language frameworks: proceedings of the ALTE Kraków Conference, July 2011* (pp. 389–404). Cambridge: Cambridge University Press.

Kostakopoulo, D. (2010). Introduction. In R. van Oers, E. Ersbøll, & D. Kostakopoulo (Eds.), *A Re-definition of belonging? Language and Integration tests in Europe* (pp. 1–23). Leiden; Boston: Martinus Nijhoff Publishers.

Krumm, H.-J. (2007). Profiles instead of levels: the CEFR and its (ab)uses in the context of migration. *The Modern Language Journal*, *91*(4), 667–669. doi:10.1111/j.1540-4781.2007.00627_6.x

Kunnan, A. J. (2009). Testing for citizenship: The U.S. Naturalization Test. *Language Assessment Quarterly*, *6*(1), 89–97. doi:10.1080/15434300802606630

Kunnan, A. J. (2012). Language assessment for immigration and citizenship. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 162–177). New York: Routlege.

L

Laversuch, I. M. (2008). Putting Germany's language tests to the test: an examination of the development, implementation and efficacy of using language proficiency tests to mediate German citizenship. *Current Issues in Language Planning*, *9*(3), 282–298. doi:10.1080/14664200802139554

Lee, J. (2010). Amendment to the Naturalization Examination and its social impact on international marriage immigrants in South Korea. *TESOL Quarterly*, *44*(3), 575–585. doi:10.5054/tq.2010.232866

Lenz, P., Andrey, S. & Lindt-Bangerter, B. (2009). *Rahmencurriculum für die sprachliche Förderung von Migrantinnen und Migranten*. Bundesamt für Migration BFM. Retrieved from http://www.bfm.admin.ch/content/dam/data/migration/integration/berichte/rahmencurriculum-d.pdf

Linn, R. L., National Council on Measurement in Education & American Council on Education (Eds.). (1989). *Educational measurement*. New York; London: Macmillan Publishers.

Little, D. (2010). The linguistic integration of adult migrants: evaluating policy and practice. Council of Europe, Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/liam/Source/Events/2010/2010evaluatingpolicy_EN.pdf

M

Mackenzie, C. (2010). Citizenship, identity, and immigration: Contemporary philosophical perspectives. In C. Slade & M. Möllering (Eds.), *From migrant to citizen: testing language, testing culture* (pp. 191–216). Basingstoke: Palgrave Macmillan.

Mar-Molinero, C. (2006). The European linguistic legacy in a global era: linguistic imperialism, Spanish and the Instituto Cervantes. In C. Mar-Molinero & P. Stevenson (Eds.), *Language ideologies, policies and practices: language and the future of Europe* (pp. 76–88). Basingstoke: Palgrave Macmillan.

Mar-Molinero, C. & Stevenson, P. (Eds.). (2006). *Language ideologies, policies and practices: language and the future of Europe*. Basingstoke: Palgrave Macmillan.

McNamara, T. (2005). 21st century Shibboleth: language tests, identity and intergroup conflict. *Language Policy*, *4*(4), 351–370. doi:10.1007/s10993-005-2886-0

McNamara, T. (2006). Validity in language testing: the challenge of Sam Messick's legacy. *Language Assessment Quarterly: An International Journal*, *3*(1), 31–51. doi:10.1207/s15434311laq0301_3

McNamara, T. (2009a). Australia: the Dictation Test redux? *Language Assessment Quarterly*, *6*(1), 106–111. doi:10.1080/15434300802606663

McNamara, T. (2009b). Language tests and social policy: A commentary. In G. Hogan-Brun, C. Mar-Molinero & P. Stevenson (Eds.), *Discourse on language and integration* (pp. 153–163). Amsterdam; Philadelphia: John Benjamins Publishing Company.

McNamara, T. (2010). The use of language tests in the service of policy: issues of validity. *Revue Française de Linguistique Appliquée*, *15*, 7–23.

McNamara, T. & Roever, C. (2006). *Language testing: the social dimension*. Malden, MA: Blackwell Publishing.

McNamara, T. & Ryan, K. (2011). Fairness versus justice in language testing: the place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, *8*(2), 161–178. doi:10.1080/15434303.2011.565438

McNamara, T. & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics*, *18*(1), 89–95. doi:10.1111/j.1473-4192.2008.00191.x

Messick, S. (1989). Validity. In R. L. Linn, National Council on Measurement in Education, & American Council on Education (Eds.), *Educational measurement* (3rd ed., pp. 13–103). New York and London: American Council on Education; Macmillan Pub. Co.; Collier Macmillan Publishers.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23. doi:10.2307/1176219

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. doi:10.1177/026553229601300302

Michalowski, I. (2010). Integration tests in Germany: a communitarian approach? In R. van Oers, E. Ersbøll & D. Kostakopoulo (Eds.), *A Re-definition of belonging? Language and integration tests in Europe* (pp. 185–210). Leiden and Boston: Martinus Nijhoff Publishers.

Michalowski, I. (2011). Required to assimilate? The content of citizenship tests in five countries. *Citizenship Studies*, *15*(6-7), 749–768. doi:10.1080/13621025.2011.600116

Milani, T. M. (2008). Language testing and citizenship: a language ideological debate in Sweden. *Language in Society*, *37*(01), 27–59. doi:10.1017/S0047404508080020

Mislevy, R. J., Almond, R. G. & Lukas, J. F. (2003). *A Brief Introduction to Evidence-Centered Design. CSE Report 632.* Los Angeles, CA: CRESST, CSE; UCLA. Retrieved from http://www.education.umd.edu/EDMS/mislevy/papers/BriefIntroECD.pdf

Mislevy, R. J., Steinberg, L. S. & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477 –496.

Möllering, M. (2010). The changing scope of german citizenship: From “guest worker” to citizen? In C. Slade & M. Möllering (Eds.), *From migrant to citizen: testing language, testing culture* (pp. 145–163). Basingstoke: Palgrave Macmillan.

Möllering, M., & Silaghi, L. (2010). From earning the priviledge of citizenship to understanding its responsibilities: An update on Australian citizenship testing. In C. Slade & M. Möllering (Eds.), *From migrant to citizen: testing language, testing culture* (pp. 236–255). Basingstoke: Palgrave Macmillan.

Müller, M., & Wertenschlag, L. (2013). „Meine Kinder möchten, dass ich auch zum Elternabend gehe“: Anmerkungen zum Szenarienansatz und zur Entstehungsgeschichte der fide-Szenarien. *Babylonia*, (1), 28–34.

N

North, B. (2009). The educational and social impact of the CEFR in Europe and beyond: a preliminary overview. In L. Taylor & C. J. Weir (Eds.), *Language testing matters: investigating the wider social and educational impact of assessment - proceedings of the ALTE Cambridge Conference April 2008* (pp. 357–378). Cambridge: Cambridge University Press.

O

Oers, R. van. (2008). From liberal to restrictive citizenship policies: the case of the Netherlands. *International Journal on Multicultural Societies*, *10*(1), 40–59.

Oers, R. van. (2010). Citizenship tests in the Netherlands, Germany and the UK. In R. van Oers, E. Ersbøll & D. Kostakopoulo (Eds.), *A Re-definition of belonging? Language and integration tests in Europe* (pp. 49–105). Leiden and Boston: Martinus Nijhoff Publishers.

Oers, R. van, Ersbøll E., & Kostakopoulo, D. (2010). Mapping the redefinition of belonging in Europe. In R. van Oers, E. Ersbøll & D. Kostakopoulo (Eds.), *A Re-definition of belonging? Language and integration tests in Europe* (pp. 307–331). Leiden and Boston: Martinus Nijhoff Publishers.

Oers, R. van, Ersbøll, E. & Kostakopoulou, D. (Eds.). (2010). *A re-definition of belonging? Language and integration tests in Europe*. Leiden and Boston: Martinus Nijhoff Publishers.

Ozolins, U. (2003). The impact of European accession upon language policy in the Baltic states. *Language Policy*, *2*(3), 217–238.

P

Papp, S. (2010). The requirements of the UK test for citizenship and settlement: critical issues and possible solutions. In L. Taylor & C. J. Weir (Eds.), *Language testing matters: investigating the wider social and educational impact of assessment - proceedings of the ALTE Cambridge Conference April 2008*. Cambridge University Press.

Perlmann-Balme, M. (2011). Deutsch-Test für Zuwanderer. Internationale Qualitätsstandards bei der Testentwicklung. *Deutsch Als Fremdsprache*, (1), 13–22.

Perlmann-Balme, M., Plassmann, S. & Zeidler, B. (2009). *Deutschtest für Zuwanderer A2-B1 Prüfungsziele, Testbeschreibung*. Berlin: Cornelsen. Retrieved from http://www.bamf.de/SharedDocs/Anlagen/DE/Downloads/Infothek/Integrationskurse/Kurstraeger/Sonstiges/dtz-handbuch_pdf

Piller, I. (2001). Naturalization language testing and its basis in ideologies of national identity and citizenship. *International Journal of Bilingualism*, *5*(3), 259–277. doi:10.1177/13670069010050030201

Plassmann, S. (2011). Aktuelle Methoden der Testmethodik und Qualitätssicherung am Beispiel des Deutsch-Tests für Zuwanderer. *Deutsch Als Fremdsprache*, (1), 23–29.

S

Saville, N. (2009). Language assessment in the management of international migration: A framework for considering the issues. *Language Assessment Quarterly*, *6*(1), 17–29. doi:10.1080/15434300802606499

Saville, N., & Van Avermaet, P. (2008). Language testing for migration and citizenship: contexts and issues. In L. B. Taylor, C. J. Weir & ALTE Conference (Eds.), *Multilingualism and assessment: achieving transparency, assuring quality, sustaining diversity: proceedings of the ALTE Berlin conference, May 2005* (pp. 265–275). Cambridge: Cambridge University Press.

Schneider, G., Neuner-Anfindsen, S., Sauter, P., Studer, T., Wertenschlag, L. & Widmer, C. (2006). *Rahmenkonzept für den Nachweis der sprachlichen Kommunikationsfähigkeit im Hinblick auf die Einbürgerung, Kurzbericht* (erstellt im Auftrag der Eidgenössischen Ausländerkommission (EKA)). Fribourg: Lern- und Forschungszentrum Fremdsprachen, Universtät Fribourg. Retrieved from http://www.ekm.admin.ch/de/dokumentation/doku/kurzbericht_rahmenkonzept.pdf

Shohamy, E. (1998). Critical language testing and beyond. *Studies In Educational Evaluation*, *24*(4), 331–345. doi:10.1016/S0191-491X(98)00020-0

Shohamy, E. (2001). *The power of tests: a critical perspective on the uses of language tests*. Harlow, England and New York: Longman.

Shohamy, E. (2006). *Language policy: hidden agendas and new approaches*. London and New York: Routledge.

Shohamy, E. (2007). Tests as power tools: looking back, looking forward. In J. Fox, M. Wesche & D. Bayliss (Eds.), *Language testing reconsidered* (pp. 141–152). Ottawa: University of Ottawa Press.

Shohamy, E. (2009). Language tests for immigrants: Why language? Why tests? Why citizenship? In G. Hogan-Brun, C. Mar-Molinero & P. Stevenson (Eds.), *Discourses on language and integration: critical perspectives on language testing regimes in Europe* (pp. 45–59). Amsterdam: John Benjamins Publishing Company.

Siiner, M. (2006). Planning language practice: a sociolinguistic analysis of language policy in post-communist Estonia. *Language Policy*, *5*(2), 161–186. doi:10.1007/s10993-006-9004-9

Skenderovic, D. (2013). Einwanderung und Sprache: Kulturalisierung einer Debatte. *Babylonia*, (1), 14–18.

Slade, C. (2010a). Civic integration in the Netherlands. In C. Slade & M. Möllering (Eds.), *From migrant to citizen: testing language, testing culture* (pp. 125–144). Basingstoke: Palgrave Macmillan.

Slade, C. (2010b). Shifting landscapes of citizenship. In C. Slade & M. Möllering (Eds.), *From migrant to citizen: testing language, testing culture* (pp. 3–23). Basingstoke: Palgrave Macmillan.

Slade, C., & Möllering, M. (2010). *From migrant to citizen: testing language, testing culture.* Houndmills, Basingstoke, Hampshire; New York: Palgrave Macmillan.

Stevenson, P., & Schanze, L. (2009). Language, migration and citizenship in Germany: discourses on integration and belonging. In G. Extra, M. Spotti & P. Van Avermaet (Eds.), *Language testing, migration and citizenship: Cross-national perspectives on integration regimes* (pp. 87–106). London and New York: Continuum International Publishing Group.

Strik, T., Böcker, A., Luiten, M. & Oers, R. van. (2010). *The INTEC Project: synthesis report. Integration and naturalisation tests: the new way to European citizenship*. Nijmegen: Centre for Migration Law, Radboud University Nijmegen.

V

Van Avermaet, P. (2009). Fortress Europe? Language policy regimes for immigration and citizenship. In G. Hogan-Brun, C. Mar-Molinero & P. Stevenson (Eds.), *Discourse on language and integration* (pp. 15–43). Amsterdam and Philadelphia: John Benjamins Publishing Company.

Van Avermaet, P. (2012). L'intégration linguistique en Europe: analyse critique. In H. Adami & V. Leclercq (Eds.), *Les migrants face aux langues des pays d'accueil: acquisition en milieu naturel et formation* (pp. 153–171). Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Van Avermaet, P. & Rocca, L. (2013). Language testing and access. In E. Galaczi & C. Weir (Eds.), *Exploring language frameworks: proceedings of the ALTE Kraków Conference, July 2011* (pp. 11–44). Cambridge: Cambridge University Press.

W

Wright, S. (Ed.). (2008). Citizenship tests in Europe - Editorial introduction. *International Journal on Multicultural Societies*, *10*(1), 1–9.

Y

Yoffe, L. (2010). Linguistic integration of adult migrants in France: policy and practice following the introduction of the reception and integration contract (CAI). *The Cultural Review -* 早稲田商学同攷会, *36*(3), 171–188.