Standards in language proficiency measurement.

> "There is a certain tendency toward standardization in the schools about which I have grave fears … Let us use standardization as a tool; let us not allow it to become a master."

(Woll, 1928)

In testing and in education, standards have been used for over a century, for the purpose of certification, for reasons of accountability, or to keep track of the performance of students and schools. The impact and prominence of standards has been rising since the 1950s (Takala et al., 2013), and by now, some educational systems have implemented language proficiency standards from the preschool level up to adult education (Cox et al., 2018). This chapter introduces three kinds language proficiency standards and critically examines their benefits and pitfalls.

DEFINING A STANDARD

We can define a standard as an agreed-upon way of doing or measuring things in order to stimulate cross-context consistency and comparability. Because of standards, screw threads are compatible worldwide and freight containers have specific dimensions. Essentially, standards are guidelines or measurable characteristics, but they have a number of specific characteristics that makes them different from regulations, which are legislative in nature (Hatto, 2010b, 2010a). In industry, standards are written by experts, reviewed by peers and ratified by a recognized body. As such, industrial standards gain legitimacy by virtue of voluntary consensus and proven utility. If a community of users fails to see the usefulness of a standard, it will receive little uptake and be rendered essentially meaningless. Generally, a standard is considered useful if it increases efficiency, quality or harmonization, or facilitates measurement. As such, a good industry standard must be precise, unambiguous, clear, and reliable within a certain margin of error (Hatto, 2010a; International Organization & for Standardization, 2016)

Using strict formatting requirements in order to promote consistency and clarity, the International Standardization Organization (ISO) has supported the development of nearly 23000 internationally recognized industry standards,

encompassing a vast array of contexts. The standards managed by ISO need to meet the key characteristics of a robust industrial standard: exactness, transparency, impartiality, effectiveness, relevance, and coherence (International Organization for Standardization, 2018).

Arguably, none of the traditional standards in language testing (e.g., CEFR, ACTFL, etc.) check all six boxes associated with a quality ISO standard. Perhaps the most common problem in language standards, is vagueness or "descriptional inadequacy" (Fulcher et al., 2011, p. 9). Quite possibly, the context-dependency and elusiveness of language proficiency (Hudson, 2012) together with the history and genealogy of the standards in use today help to explain why this is the case. To better understand the standards used in language performance measurement today, this chapter takes a broad and historical perspective.

TYPES OF LANGUAGE PROFICIENCY STANDARDS

In this chapter, we distinguish between three types of standards: educational performance indicators (e.g., PISA, PIRLS, NAEP, PIAAC), language proficiency frameworks (e.g., ACTFL, CEFR), and institutionalized language tests (e.g., IELTS, TOEFL). These three types of standards have received wide and often voluntary uptake by a community of users, but serve dissimilar goals.

The goal of educational performance indicators is to compare the performance of L2 learners or L1 users by administering the same test in different educational settings. These assessment-driven standards typically focus on a range of domains (e.g., science, mathematics), but often include a linguistic component. Educational performance indicators allow policy makers and researchers to monitor the performance of their own educational system over time, but also trace its international ranking in relation to the educational systems in other states or countries. As such, they aim to impact policy or spur policy change (Singer et al., 2018). International educational performance assessments reside under the auspices of organizations such as IEA (International Association for the Evaluation of Educational Achievement) or OECD (Organization for Economic Cooperation and Development) and often rely on international research teams to conduct and analyze the assessments.

Centralized tests are at the heart of educational performance indicators; the performance levels and thresholds are inextricably tied to one specific measurement

instrument. The same is not the case for language proficiency frameworks, in which the level descriptors take center stage. They serve to offer an external benchmark of language proficiency for examination boards, policy makers, test takers and other stakeholders to refer to (Fulcher, 2016).

Some organizations developing language proficiency frameworks (e.g., the performance descriptors by the *American Council On The Teaching Of Foreign Languages*, or ACTFL) do also offer tests, but the frameworks are not meaningless without them – as they would be in the case of a PISA or a PIRLS test. Language proficiency frameworks provide generic descriptions of different levels of L2 proficiency. Often, they are language-independent (e.g., the Common European Framework of Reference for Languages, or CEFR, Council of Europe, 2001), but some are dedicated to a specific language (e.g. CEFR-J, an adaptation of the CEFR for Japanese learners of English. See Negishi & Tono, 2016), or a limited number of languages (e.g. *Canadian Language Benchmarks*, or CLB, for English and French. See Centre for Canadian Language Benchmarks, 2012). Often, the descriptors are context-independent, but some language proficiency frameworks are quite context-specific indeed (e.g., STANAG 6001. See NATO, 2014). Many, but not all, language proficiency frameworks have ties to national or supra-national politics (e.g., *China Standards of English*. See Ministry of Education of the People's Republic of China, 2018), while others are private nonprofit organizations (e.g., ACTFL).

With Hatto (2010b), standards issued by national or international bodies but created by commissions of independent experts could be called *formal standards*. *Private standards* on the other hand, are created by private exam boards such as IELTS or TOEFL. The band levels developed by these exam boards might not have been intended as standards, but have started being used as such due to their widespread use for high-stakes purposes.

THE POLICY PERSPECTIVE: EDUCATIONAL PERFORMANCE INDICATORS

The roots of the largest international educational performance indicators such as PIRLS (*Progress of International Reading Literacy Study*, see: Mullis, Martin, Foy, & Hooper, 2017), PISA (*Programme for International Students Assessment*, see: OECD, 2016), and PIAAC (*Programme for the International Assessment of Adult*

Competencies, see: OECD, 2019) can be traced back to the 1950s, but arguably, they have never been quite as prominent as they are now.

Educational performance indicators are focused on measuring and comparing the (reading) performances of schools, educational systems, and economies (Takala et al., 2013). Their aim is not to certify but to chart trends and to offer data that can impact policy (Singer et al., 2018). Most people who use the data or read the reports will thus adopt a norm-referenced approach, scrutinizing the relative position of one country's educational performance over time, or in relation to other systems at the same point in time. Additionally, however, international educational performance indicators have a criterion-referenced component, which can be interpreted as minimal performance standards. In PISA, for example, Level 2 is also called the baseline level, implying that 15-year olds should be able to perform tasks at this level in order to be able to participate in society (see Table 1).

Table 1. PISA reading performance levels

Level 6 Making multiple inferences, comparisons and contrasts that are both detailed and precise.
Level 5 Locating and organizing several pieces of deeply embedded information, inferring which information in the text is relevant.
Level 4 Locating and organizing several pieces of embedded information or interpreting the meaning of nuances of language in a section of text by taking into account the text as a whole.
Level 3 Recognizing the relationship among several pieces of information, or integrating several parts of a text in order to identify a main idea.
Level 2: baseline level Retrieving one or more pieces of information that may have to be inferred.
Level 1a Retrieving one or more independent pieces of explicitly stated information, or interpreting the theme or purpose of a text on a familiar topic.
Level 1b

Retrieving a single piece of explicitly stated information in a short, syntactically simple text with a familiar context and text type.

Note. The original – substantially longer and more detailed – descriptors have been edited from OECD, 2016, pp. 164–166.

It is important to stress that the organizations behind these initiatives have different agendas and goals, which may impact the construct of the tests. IEA, which develops PIRLS, is a conglomerate of research centers, government-affiliated research institutions, researchers and analysts. Their stated goal is to "research, understand, and improve education worldwide" (IEA, 2019). In line with this mission, PIRLS aims "to provide the best policy-relevant information about how to improve teaching and learning and to help young students become accomplished and self-sufficient readers" (Mullis et al., 2017, p. 4). The OECD, which funds and – through various local research partners and subcontractors – runs PISA and PIAAC, is an economic partnership between 36 mainly Western member states, which aims "to shape policies that foster prosperity, equality, opportunity and well-being for all" and "to better prepare the world of tomorrow". Correspondingly, the goals of PISA and PIAAC are focused on participation in society and in the labor market (OECD, 2016, 2019).

In spite of – or because of – their impact, educational performance indicators have come under intense scrutiny in recent years. PISA especially has been challenged on a number of fronts. Perhaps the most widely supported condemnation of PISA to date was published in 2014, when over eighty academics and educationalists called for a moratorium on the test (The Guardian, 2014). The main strands of criticism on OECD-based international educational performance indicators focus on their impact and their construct. Critics argue that the relatively short three-year cycles of these assessments encourage policy makers to think in terms of short-term fixes (The Guardian, 2014). Related to this, OECD assessments have been criticized for not being attuned to local educational systems (Gaber et al., 2012; Liss, 2013; Sjøberg, 2015), for being restricted to what can easily be measured, and for reducing the aim of education to preparation for the labor market (Sjøberg, 2015; The Guardian, 2014; Y. Zhao, 2020).

THE CAN-DO PERSPECTIVE: LANGUAGE PROFICIENCY FRAMEWORKS

The Foundational years (1950s-1970s)

A pioneer in the field of language proficiency standards, Kaulfers was one of the first people to observe that that test scores do not carry meaning in and of themselves. Because the same score on different tests might say something very differently about a test taker's ability, Kaulfers (1944) introduced a listening proficiency scale, linked to a listening test. Table 2 lists Kaulfers' level descriptors, which – if only for the can-do approach – remind of listening descriptors still in use.

Table 2. Kaulfers' aural performance Scale (Kaulfers, 1944, p. 139)

Level	Descriptor
4	Can understand popular radio talks, talking-pictures, ordinary telephone conversations, and minor dialectal variations without obvious difficulty.
3	Can understand ordinary conversation on common, topics, with the aid of occasional repetition or paraphrastic statements.
2	Can understand the ordinary questions and answers routine transactions involved in independent travel.
1	Can catch a word here and there and occasionally guess meaning through inference.
0	Cannot understand the spoken language.

Note. Level number added by author

Innovations in educational measurement often result from a societal or institutional need to achieve a greater degree of order (Stein, 2016). And so, when in 1952 the US Civil Service Commission wanted to create an inventory of the foreign language proficiency of government employees and found that there was little or no consistency or system in dizzying array of incompatible scores and profiles reported, it was decided to create a single L2 proficiency scale that all employees could refer to (Sollenberger, 1978; Spolsky, 1995).

In his first-hand account, Sollenberger (1978) explains that the FSI scale (Foreign Service Institute) – as it was to be known – was to comprise of six levels,

ranging from zero (no competence) to five (native speaker or bilingual competence) but there was no clear theoretical or empirical rationale for choosing a six-level scale. In 1955, based on the outcomes of a pilot of the FSI scale among two hundred officers, the descriptors were refined and subdivided into speaking and reading scales. The first FSI definitions of speaking proficiency levels were adopted in 1956, with John B. Carroll consulting on the test construction and scale revision process (Liskin-Gasparro, 1984).

Carroll approved of the scaling efforts made by Kaulfers, but resisted the idea of conceptualizing language skills as unitary traits. As such, he supplemented the general speaking scale with more specific descriptors: accent, grammar, vocabulary, fluency, and comprehension (Carroll, 1954). For each of these scales, six proficiency levels were defined. In 1968, these level descriptors were revised, standardized (on educated native speakers) and adopted by the Interagency Language Roundtable (ILR) (Liskin-Gasparro, 1984). The first part of each ILR level descriptor is displayed in Table 3. All levels except for level 5 could be modified by a "+" to indicate a performance that exceeds the minimum requirements of one level but falls short of the next.

Table 3. FSI/ILR speaking descriptors (1968)

Level	Code	Definition
Native or Bilingual Proficiency	S-5	Speaking proficiency equivalent to that of an educated native speaker. Has complete fluency in the language such that his speech on all levels is fully accepted by educated native speakers …
Full Professional Proficiency	S-4	Able to use the language fluently and accurately on all levels normally pertinent to professional needs. Can understand and participate in any conversation within the range of his experience …
Minimum Professional Proficiency	S-3	Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics. Can discuss particular interests …

Limited working proficiency	S-2	Able to satisfy routine social demands and limited work requirements. Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events …
Elementary Proficiency	S-1	Able to satisfy routine travel needs and minimum courtesy requirements. Can ask and answer questions on topics very familiar… can understand simple questions and statements, allowing for slowed speech …
No proficiency	S-0	

Note. Descriptors trimmed to first thirty words. (For full descriptors, see Sollenberger, 1978, pp. 19–22)

It is difficult to overstate the importance of the pioneering work done on the FSI scales, which directly influenced many of the standards in use today in form and content. In a modified format the FSI scales are still in use and are known as ILR scales, as they resulted from a 1985 Interagency Language Roundtable that revised the scales and formally included the plus levels.

STANAG 6001: The ILR scale goes transatlantic

In the 1970s the Bureau for International Language Coordination (BILC) started adapting the ILR descriptors for use among NATO member states. When NATO accepted the standards in 1976 they became known as Standardization Agreement 6001 (STANAG 6001). The purpose of the standards was to facilitate international staff recruitment and to be able to compare national language proficiency standards by using one and the same descriptive system (Green & Wall, 2005). The standards are available for English and French – the two working languages of NATO.

Some fifty years after the launch of STANAG 6001, some fundamental problems still prevent it from being a vehicle for full comparability of language proficiency levels across NATO member states. A major issue has to do with interpretability (Brooks & Hoffman, 2013; Dubeau, 2006). Much like the CEFR or the ILR descriptors, the wording in the STANAG 6001 descriptors may be somewhat vague and may leave room for interpretation. Additionally, because they include both general purpose and specifically military descriptions of language proficiency the

STANAG 6001 descriptors may not always be entirely suited to the specifics of military communication (Brooks & Hoffman, 2013; Fulcher, 2015). Other problems that may impede the optimal operationalization of the STANAG 6001 descriptors in the field include a lack of theoretical basis in second language acquisition theory, and somewhat vague level demarcations (Green & Wall, 2005). In order to achieve a more uniform interpretation of the STANAG 6001 descriptors, BILC members continue to have regular standardization meetings. Additionally, every three years the STANAG 6001 descriptors are under mandatory revision (NATO, 2014). Currently the fifth major revision is in use.

ACTFL: An academic language proficiency framework

For the first decades following the introduction of language proficiency standards, academic interest was limited, so the first standardization efforts and language proficiency frameworks resulted from government or military initiatives (Spolsky, 1995). By the mid-to-late 1970s, the interest in proficiency guidelines and curricula started growing in academic circles (Green & Wall, 2005), and in the 1980s and 1990s a growing concern for accountability helped spur the creation and proliferation of the ACTFL guidelines (Cox et al., 2018; Ricardo-Osorio, 2008). In 1982, ETS and ACTFL had begun work on the development of provisional proficiency guidelines that would translate the ILR scale to an academic context (Brooks & Hoffman, 2013; Liskin-Gasparro, 1984. For a detailed overview of ACTFL's history, see Cox et al., 2018). Table 4 shows how the current ACTFL levels relate to the original ILR levels.

The ACTFL guidelines describe what learners can do at specific proficiency levels. The scale identifies eleven language proficiency levels that encompass five main levels (from novice to distinguished) and three sublevels (Low, Mid, High). Table 4 also shows these levels and their CEFR counterpart.

Table 4. ACTFL alignment with ILR scales and CEFR levels

ILR – ACTFL concordance[1]	ACTFL – CEFR concordance[2]

	Receptive skills		Productive skills	
ILR	ACTFL	CEFR	ACTFL	CEFR
4-5	Distinguished	C2		
3-4	Superior	C1.2	Superior	C2
2+	Advanced High	C1.1	Advanced High	C1
2	Advanced Mid	B2	Advanced Mid	B2.2
2	Advanced Low	B1.2	Advanced Low	B2.1
1+	Intermediate High	B1.1	Intermediate High	B1.2
1	Intermediate Mid	A2	Intermediate Mid	B1.1
1	Intermediate Low	A1.2	Intermediate Low	A2
0+	Novice High	A1.1	Novice High	A1
	Novice Mid	0	Novice Mid	0
	Novice Low	0	Novice Low	0

Note.

1. See Brooks & Hoffman (2013); Hudson (2012); Liskin-Gasparro (1984)

2. See ACTFL (2016)

What sets ACTFL apart from other language proficiency frameworks is that it does not only produce guidelines and descriptors, but it also controls ACTFL-based assessment through Language Testing International (LTI)[1]. Another defining feature of ACTFL is its formal integration in the US educational system (Cox et al., 2018, p. 105). Having been incorporated in the *National Standards for Foreign Language Learning in the 21st Century* ACFTL now impacts classroom practice in the US from K-12 to college (Glisan, 2012; Little, 2019).

Like any successful language proficiency framework that has gained global resonance, the ACTFL guidelines have also been subject to criticism. Notably, Fulcher (1996) criticized them for lacking an empirical foundation, for disregarding the qualities and imperfections of real-world utterances and for an overreliance on native speaker norms (for a full discussion, see Liskin-Gasparro, 2003). Subsequent ACTFL revisions (ACTFL, 2012) have tried to answer these points of criticism, but according to some reviewers they still apply to a certain degree (Little, 2019) .

[1] For STANAG 6001 there is the centralized *SHAPE* (Supreme Headquarters Allied Powers) test, but individual NATO member states are free to develop and administer their own STANAG 6001-based test alongside *SHAPE*.

The CEFR: Policy meets practice

ACTFL and the CEFR are sometimes seen as two transatlantic sides of the same coin, but there are a number of important differences. The CEFR is embedded in the Council of Europe's vision of multilingualism and lifelong learning (Council of Europe, 2001, 2018; Trim, 2012) and is focused on language learning and teaching. ACTFL's prime focus is on assessment (see Little, 2019). The CEFR is essentially a language policy document, and quite possibly the most influential language policy document to date (Figueras, 2012).

Its roots can be traced back to the 1970s, when the Council of Europe's efforts to examine the possibility of creating a unified system of expressing language proficiency levels led to *The Threshold Level* (van Ek, 1975; van Ek & Trim, 1991a). This level would later be known as B1, and described the linguistic context and challenges of L2 users living in a foreign European country. In the wake of *The Threshold Level's* success, the new language proficiency levels *Waystage* (van Ek & Trim, 1991b) and *Vantage* (van Ek & Trim, 2001) were described and introduced. When the CEFR was published in 2001, it incorporated the work done on the *Threshold*, *Vantage* and *Waystage* levels, as well as numerous other language proficiency scales, including the ILR levels (Council of Europe, 2001).

Table 5 shows the structure underlying the six CEFR levels: three broad levels identify basic, independent and proficient language users, and within each of these three levels, two sublevels exist. In some scales these sublevels are further subdivided into a basic level and a so-called "plus-level".

Table 5. Organization of CEFR levels

Purpose[1]	Level		User categorization
Survival	A1	Breakthrough	Basic user
Routine transaction & interaction	A2	Waystage	
	B1	Threshold	Independent user
Academic / Professional	B2	Vantage	
	C1	Effective Operational Proficiency	Proficient user
	C2	Mastery	

[1] See Little (2019)

CEFR level descriptors are meant to be general, language-independent and context-independent descriptions of what learners can do with language (Hudson, 2012; North, 2014a). The impressionistic nature of the CEFR level descriptor means that test developers using the framework are required to adapt it to their own local context and to the needs of the language learners at hand (North, 2014b). It also means that the CEFR allows for – encourages – the development of differentiated profiles or requirements (e.g., uneven requirements, such as writing at B1, speaking at B2).

Of all of the content the CEFR offers, the level descriptors have drawn most of the attention. More than likely, the level descriptors, listed in tables, are what have given the CEFR such wide appeal among teachers, testers, publishers and policy makers (Figueras, 2012; Little, 2007; Trim, 2012). The scholars who built the CEFR's foundation were initially hesitant of compartmentalizing language proficiency in levels, organized in hierarchical tables, but decided to use that structure for practical reasons. As John Trim recalled, they had not anticipated the appeal of this approach to curriculum designers and policy makers:

> "*Practical considerations overrode the theoretical misgivings as to the validity if the concept of 'level'. We had used the term 'level' originally despite deep misgivings about the concept. We could see no reason to break the process of language learning into a series of steps […] Over time, it became apparent*

that our reasoning took too little account of the realities of the social organization of language learning."(Trim, 2012, p. 29)

The CEFR was always intended to be more than a collection of illustrative tables (North, 2014a): At its heart are the values of multilingualism, lifelong learning, and the international mobility of people and ideas. However, over recent years it has become clear that the CEFR is also being used for purposes inimical to these original ideals. Normative use of the illustrative descriptors has been frequently observed (Deygers et al., 2018; Fulcher, 2004), and CEFR levels are being used to justify migration policies meant to curb the movement of people and to support monolingual ideologies (Rocca et al., 2019). These misuses of the CEFR have engendered a strand of CEFR criticism that is focused on its political use (Barni, 2015). Other authors have criticized the methodological foundations of the framework (Alderson, 2007), the lack of theoretical support from SLA theory (Hulstijn, 2007) and the vagueness of the level descriptors (Galaczi et al., 2011). The recently published CEFR *Companion Volume* (Council of Europe, 2018) answers some of the critique on the original CEFR: it no longer upholds a native speaker norm (see McNamara, 2014), pays more attention to multilingualism (Krumm, 2007) and focuses on the concept of mediation. Many of the most fundamental critiques on the CEFR have not been addressed in the Companion Volume, however (Deygers, 2019).

In spite of the criticism, the CEFR has been a remarkably successful document that has received global and voluntary uptake in a wide array of contexts and applications. It has transformed language teaching, language testing and language policy in Europe and around the world, and has inspired the creation of new language proficiency frameworks adapted to local contexts, such as the CEFR-J (Negishi & Tono, 2016), which describes the English language proficiency of Japanese L1 learners.

Canadian Language Benchmarks: Task-based can-dos

The realization that test scores provided by different language schools and exam boards often lack transparency is what started the FSI initiative, what inspired the creation of the CEFR and what led BILC to start work on STANAG 6001. It is also

what led the Canadian government to fund the development of the Canadian Language Benchmarks (CLB) in the 1990s (Peirce & Stewart, 1997).

Rooted in Bachman and Palmer's theory of communicative language ability, the CLB provide a descriptive continuum of language ability that is competency-based, task-based, and learner-centred (Centre for Canadian Language Benchmarks, 2012). The CLB offer descriptive statements of twelve language proficiency levels for the four traditional language skills in specified performance conditions. The Benchmarks describe language tasks that learners should be able to perform at these levels. Additionally, the CLB function as a national standard for curriculum planning so as to streamline the ESL tuition for migrants across Canada.

Table 6 shows how the twelve benchmarks are structured, and how they relate to the CEFR (North & Piccardo, 2018). The collaboration between the CEFR and CLB to align two frameworks reflects the similarities between the two frameworks, which share a can-do approach, a focus on communicative competence, and a learner-oriented philosophy.

Table 6. Overview of the CLB and their relationship with the CEFR levels

Basic Language Ability		
Communicating in predictable context about basic needs and familiar matters		
CLB1	Initial	Pre-A1
CLB2	Developing	A1
CLB3	Adequate	A1/A2
CLB4	Fluent	A2
Intermediate Language Ability		
Functioning independently in familiar situations and a few less well-known contexts		
CLB5	Initial	B1
CLB6	Developing	B1
CLB7	Adequate	B1
CLB8	Fluent	B2
Advanced Language Ability		
Effective, accurate and appropriate communication in a wide range of contexts.		
CLB9	Initial	B2
CLB10	Developing	C1

CLB11	Adequate	C1
CLB12	Fluent	C2

China Standards of English: a curriculum-based standard

In an endeavor to establish a common standard of English, various Asian countries have adopted (Taiwan, see Wu, 2014) or adapted (Japan, see Negishi & Tono, 2016) the CEFR. The Chinese government, however, decided to create a language framework specific to the Chinese context (Jin et al., 2017). The China Standards of English (CSE) are meant to streamline the English language curriculum in China from elementary school to university and to create shared performance standards for teaching, learning and assessment. The CEFR inspired but did not guide the development of the China Standards of English (Jin et al., 2017). Nevertheless, even though the CEFR is not mentioned in the CSE, it clearly inspired the can-do approach, the organization of the document, and the use of an overall scale in combination with self-assessment scales and subscales (Ministry of Education of the People's Republic of China, 2018).

The CSE define nine levels of English language proficiency, organized in three stages. In line with the CSE goals, the levels have been linked to the curriculum. Table 7 displays the organization of these levels, shows how the CSE vocabulary standards have been calibrated against the CEFR, and links the levels to the Chinese curriculum (W. Zhao et al., 2017).

Table 7. CSE overview and CEFR link

CSE stage	CSE level	CEFR[1]	Curriculum
Advanced stage	Level 9		
	Level 8	C1	Advanced College English Curriculum Requirement
	Level 7	B2	Intermediate College English Curriculum Requirement
Intermediate stage	Level 6	B1+	Basic College English Curriculum Requirement
	Level 5	B1	End of senior secondary school

	Level 4		
Elementary stage	Level 3	A2	End of junior secondary school
	Level 2		
	Level 1	A1	End of primary school

[1] The CEFR alignment applies specifically to vocabulary, see W. Zhao et al. (2017)

THE CERTIFICATION PERSPECTIVE: LARGE-SCALE LANGUAGE TESTS

In this section we focus on test scores that are used as standards. As a clear case in point, we will use the context of university admission, where IELTS and TOEFL have come to be seen as standards because they have gained wide acceptance (Hyatt & Brooks, 2009).

In the English-speaking world, TOEFL and IELTS have traditionally dominated the competitive market of language assessment for university admission, and are still the most widely accepted tests (Green, 2018; Weigle & Malone, 2016). These institutionalized high-stakes international language tests are clearly distinct from international educational performance indicators, because they do not aim to impact educational policies but focus on certifying individual performances. In fact, *Educational Testing Systems* (ETS), the organization behind TOEFL, explicitly disapproves of the practice of ranking countries on the basis of TOEFL scores and regards this as "a misuse of data" (ETS, 2018, p. 13). Being international language tests, TOEFL and IELTS are also different from language proficiency frameworks since their aim is not to describe but to evaluate language proficiency (IELTS, 2019). The fact that certain scores on these tests are being used as de facto standards is not a policy that is pursued by the test developers.

The University of Cambridge Local Examinations Syndicate (UCLES) started developing and administering English language tests for non-native speakers of English in 1913 (Weir, 2003). TOEFL was first introduced some fifty years later, by the National Council on the Testing of English as a Foreign Language (Spolsky, 1995). IELTS was established in the 1980s (Davies, 2008) and has grown to become the largest

international English language test worldwide. It was taken by 3.5 million test takers in 2018 (British Council, 2019a)[2].

These tests focus on scoring and certification but also offer test taker profiles at specific score levels. TOEFL candidates receive separate scores in reading, listening, writing and speaking. Each skill is scored on a thirty-point scale, and performance descriptors have been drawn up for four score ranges on this scale, ranging from "Below Low- Intermediate" to "Advanced" (ETS, 2019). An advanced reader, for example, will typically "understand academic passages in English at the introductory university level […]" (ETS, 2019a, p. 1). IELTS uses a system of ten bands (0-9) to assess reading, listening, writing and speaking. The four scores on the separate skills are then averaged to yield a final band score. Six profiles have been linked to these overall band scores, ranging from "Limited User" to "Expert User" (see Table 3). A Good user (i.e. band score 7) is described as somebody with "an operational command of the language, though with occasional inaccuracies, inappropriate usage and misunderstandings in some situations (British Council, 2019b)".

In university admission requirements scores of 6.5-7.5 on IELTS and a TOEFL iBT score range of 90-110 have come to be seen as sufficient proof of a language proficiency level that is minimally required to commence academic studies in English. Table 8 shows the minimum IELTS and TOEFL requirements for international students to the ten highest-ranking universities in the Times Higher Education index.

Table 8. Minimum English language proficiency requirements for international students in arts and humanities

	TOEFL iBT	IELTS
Stanford University, (2019)	100	Not accepted
The University of Cambridge (2017)	110	7.5
University of Oxford (2019)	110	7.5
Massachusetts Institute of Technology (2019)	90	7
Harvard University (2019)	English test not required	
University College London (2018)	100	7
Princeton University (2019)	Level requirement not stated	
The University of Chicago (2019)	100	7

[2] Over the years, TOEFL has been administered to over 35 million test takers. The organization does not wish to share or publish annual numbers of test takers, however (private communication, 22 July 2020).

Yale College (2019)	100	7
University of California, Berkeley (2019)	80	6.5

Note. Selected universities represent the top ten universities of the Times Higher Education ranking for Arts & Humanities

In the Anglo-Saxon university admission policy, language test scores are more often used than language proficiency framework levels (Deygers et al., 2018). Since not all institutions make use of language proficiency frameworks, language test providers have drawn up score equivalence guidelines to help institutions with setting level requirements. Table 9 shows two such score equivalence tables; one equating TOEFL iBT scores with IELTS bands, published by ETS (2010), and one published by Pearson (2019), equating PTE Academic scores with IELTS and TOEFL scores. As the table shows, the outcomes of the two equivalence tables are not quite the same. For example: the ETS study finds that an overall IELTS score of 5.5 corresponds to a TOEFL iBT 46-49, but the Pearson study equates the same IELTS score with a TOEFL iBT score range of 54-56. The same dissimilarities can be found in the right-hand part of table 9. When plotting the CEFR score alignment outcomes of the three tests (De Jong et al., 2014; ETS, 2010; Riazi, 2013; Taylor, 2004) it is easy to see that the threshold levels for the different CEFR levels do not quite match (for an extensive study on this topic, see Green, 2018).

Table 9. Score equivalence guides compared

ETS score equivalence[1]		Pearson score equivalence[2]			CEFR alignment		
IELTS	TOEFL iBT	IELTS	PTE A	TOEFL iBT	IELTS[3]	TOEFL iBT[1]	PTE A[4]
0-4	0-31						
4					B1		
4.5	32-24	4.5	30			B1	A2
5	35-45	5	36			(24-35)	(30-42)
			38	40-44			
5.5	46-59	5.5	42	54-56	B2	B2	B1
			46	65-66		(46-93)	(43-58)
		6	50	74-75			

6	60-78		51				
			53	79-80			
6.5	79-93	6.5	58		C1		
			59	87-88			B2
			64	94		C1	(59-75)
7	94-101	7	65			(94-114)	
			68	99-100			
			72	105			
7.5	102-109	7.5	73		C2		
			78	113			C1
8	110-114	8	79			C2 (115-120)	(76)
8.5	115-117	8.5	83				
			84	120			
9	118-120	9	86				

[1] (ETS, 2010)

[2] (Pearson, 2019a)

[3] (Taylor, 2004)

[4] (Riazi, 2013)

Like most language proficiency standards, TOEFL and IELTS have been subjected to criticism. One strand of criticism focuses on the use of tests scores. A wide range of studies have documented consistent uneducated or unsubstantiated use of test scores for university admission (Baker, 2014; Deygers & Malone, 2019; Hyatt, 2013; O'Loughlin, 2008, 2013). Setting score thresholds in a simplistic manner without a careful analysis goes against the recommendations of the test providers, undercuts a usage-based validity argument in the line of Kane (2013) or Bachman & Palmer (2010), and undermines the credibility of test scores as standards of academic language proficiency. Other researchers have criticized the power that these tests have over the lives of individuals (Hamid & Hoang, 2019) and the lack of accountability they face for the scores they assign (Pearson, 2019b; Sarich, 2012). However, given the rise of international student mobility (Cantwell, 2015; UNESCO, 2018) and the increasing need for documented English language proficiency in the context of education, employment and migration, it is unlikely that major high-stakes test of English will soon see a reduction in terms of candidature, or that their levels will cease to be used as language proficiency standards in certain contexts.

CONCLUSON: ARE LANGUAGE PROFICIENCY STANDARDS, STANDARDS?

This chapter has provided an overview of three types of language proficiency standards, and discussed a range of concrete examples (PIRLS, PISA, PIAAC, the FSI/ILR scales, STANAG 6001, the ACTFL guidelines, the CEFR, the Canadian Language Benchmarks, the China Standards of English, IELTS, and TOEFL). Many of these language proficiency standards are similar but not the same. Many are somehow linked or aligned, but not always robustly so. All standards discussed in this chapter have been widely adopted, but none have been free of fundamental criticism. Commonly recurring critiques include construct irrelevance, theoretical underspecification, and descriptive vagueness. Educational performance indicators such as PISA or PIRLS are intended to be used on a population level and have proven to be highly influential in guiding educational policies. Critics have argued that OECD-based standards focus too heavily on education in function of the economy and do not take into account the variety of educational practices across the globe. The use of large-scale language tests as de facto standards in high-stakes contexts (e.g., university admission, citizenship procedures) has become widespread, but is not uncontested. Uneducated use of test scores has been widely documented and has been the subject of criticism.

It is unlikely that any of the language proficiency standards discussed in this chapter meet the defining criteria of a standard: specificity, reliability and universal recognition (Hatto, 2010b). Of course standardizing screw threads and freight containers is a more clearly delineated and concrete practice than standardizing levels of language proficiency (see also Fulcher, 2016). Authors have argued that language is changeable across contexts, that each of the world's 7000 languages has its own unique lexico-grammatical system, that every speaker has their own idiolect, and so on (Hudson, 2012). There is truth to this argumentation, and it helps to explain the diversity of standards, but it does not adequately explain why language and context-specific standards (e.g., STANAG 6001) suffer from the same theoretical and descriptive shortcomings as general, language-independent standards (e.g., CEFR).

Variability alone cannot account for the descriptional inadequacy that typifies many language performance standards today. Their shortcomings can probably also be explained – at least in part – by the fact that many build on intuitions and descriptions

going back to the 1950s, on reified language proficiency levels (Fulcher, 2004), or on descriptors that may have a rather thin empirical basis. As such, to an extent, the broad field of language assessment continues use standards that are supported by shaky foundations (Hulstijn, 2007) and explains the shortcomings of these standards by invoking the elusive variability of language.

To achieve greater robustness and less variability we may need to fundamentally rethink the standards we currently use. Such a reconceptualization is drastic but not impossible; outside of linguistics too, widely used standards have been changed when the need for improvement was apparent. In 2018 the International Bureau of Weights and Measurement decided to change the definition of the kilogram, the ampere, the kelvin and the mole by linking them all to the same seven constants. It is not impossible to conceive a similar shift taking place in language measurement, if language proficiency standards are to truly function as veritable standards. A truly empirical, performance-driven description of language could provide a good starting-point (Fulcher, 2012). Quite possibly, natural language processing will be instrumental in creating language performance standards that are able to achieve greater levels of precision.

SUGGESTED FURTHER READING

Stein, Z. (2016). *Social Justice and Educational Measurement*. Oxon and New York: Routledge. This original and insightful book applies ideas from moral and political philosophy to educational assessment. Stein argues that standardized assessment has not led to increased educational equality and proposes a model that offers a solution for how standardized testing could be used to foster equal educational opportunities.

Sollenberger, H. E. (1978). Development and current use of the FSI Oral Interview Test. In J. L. D. Clark (Ed.), *Direct Testing of Speaking Proficiency: Theory and Application*. (pp. 1–13). Princeton, N.J.: Educational Testing Service. When tracing the history language performance standards it is worth going back to first-hand accounts of the development of the precursors to today's standards.

Trim, J. L. M. (2012). The Common European Framework of Reference for Languages and its background: A case study of cultural politics and educational

influences. In M. Byram & L. Parmenter (Eds.), *The Common European Framework of Reference: The Globalisation of Language Education Policy* (pp. 14–36). Buffalo: Multilingual Matters. John Trim, one of the people who laid the groundwork for the CEFR, remembers which ideas, intentions and aspirations inspired the creation of what was to become perhaps the most widely used set of language performance standards.

References

ACTFL. (2012). *ACTFL Proficiency Guidelines 2012*. American Council on The Teaching of Foreign Languages.

ACTFL. (2016). *Assigning CEFR Ratings to ACTFL Assessments*. American Council On The Teaching Of Foreign Languages. https://www.actfl.org/publications/guidelines-and-manuals/assigning-cefr-ratings-actfl-assessments

Alderson, J. C. (2007). The CEFR and the Need for More Research. *The Modern Language Journal*, *91*(4), 659–663. https://doi.org/10.1111/j.1540-4781.2007.00627_4.x

Bachman, L., & Palmer, A. (2010). *Language Assessment in Practice*. Oxford University Press, USA.

Baker, B. A. (2014). *Investigating language assessment literacy in canadian university admissions*. LTRC 2014, Amsterdam.

Barni, M. (2015). In the name of the CEFR: Individuals and Standards. In B. Spolsky, O. Inbar-Lourie, & M. Tannenbaum (Eds.), *Challenges of language education and policy. Making space for people* (pp. 40–52). Routledge.

British Council. (2019a). *IELTS grows to 3.5 million a year*. Take IELTS. https://takeielts.britishcouncil.org/about/press/ielts-grows-three-half-million-year

British Council. (2019b). *Understand and explain the IELTS scores*. Take IELTS. https://takeielts.britishcouncil.org/teach-ielts/test-information/scores-explained

Brooks, R. L., & Hoffman, M. (2013). Government and Military Assessment. In *The Companion to Language Assessment*. John Wiley & Sons, Inc. http://onlinelibrary.wiley.com/doi/10.1002/9781118411360.wbcla069/abstract

Cantwell, B. (2015). Are international students cash cows? Examining the relationship between new international undergraduate enrollments and institutional revenue at public colleges and universities in the US. *Journal of International Students*, *5*(4), 512–525.

Carroll, J. B. (1954). *Notes on the Measurement of Achievement in Foreign Languages*.

Centre for Canadian Language Benchmarks. (2012). *Canadian Language Benchmarks. English as a second language for adults*. Citizenship and Immigration Canada.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Council of Europe.

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors*. Council of Europe Language Policy Division.

Cox, T. L., Malone, M. E., & Winke, P. (2018). Future directions in assessment: Influences of standards and implications for language learning. *Foreign Language Annals*, *51*(1), 104–115. https://doi.org/10.1111/flan.12326

Davies, A. (2008). *Assessing academic English: Testing English proficiency 1950 - 89: the IELTS solution*. Cambridge Univ. Press.

De Jong, J. H. A. L., Becker, K., Bolt, D., & Goodman, J. (2014). *Aligning PTE Academic Test Scores to the Common European Framework of Reference for Languages*. Pearson. https://pearsonpte.com/wp-content/uploads/2014/07/Aligning_PTEA_Scores_CEF.pdf

Deygers, B. (2019). The CEFR Companion Volume: Between Research-Based Policy and Policy-Based Research. *Applied Linguistics*. https://doi.org/10.1093/applin/amz024

Deygers, B., & Malone, M. E. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*, *36*(3), 347–368. https://doi.org/10.1177/0265532219826390

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One Framework to Unite Them All? Use of the CEFR in European University Entrance Policies. *Language Assessment Quarterly*, *15*(1), 3–15. https://doi.org/10.1080/15434303.2016.1261350

Dubeau, J. (2006). *Are We All On the Same Page? An Exploratory Study of OPI Ratings across NATO Countries Using the NATO STANAG 6001 Scale*.

ETS. (2010). *Linking TOEFL iBT TM Scores to IELTS® Scores – A Research Report*. Educational Testing Service.

ETS. (2018). *Test and score data summary for TOEFL iBT® tests. January 2017–December 2017*. Educational Testing Service.

ETS. (2019). *Performance descriptors for the TOEFL iBT® test*. Educational Testing Service.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, *66*(4), 477–485. https://doi.org/10.1093/elt/ccs037

Fulcher, G. (1996). Invalidating validity claims for the ACTFL Oral Rating Scale. *System*, *24*(2), 163–172.

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

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 378–392). Routledge.

Fulcher, G. (2015). *Re-examining Language Testing: A Philosophical and Social Inquiry*. Routledge.

Fulcher, G. (2016). Standards and frameworks. In D. Tsagari & J. Banerjee (Eds.), *Handbook of Second Language Assessment* (pp. 29–44). De Gruyter Mouton.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29. https://doi.org/10.1177/0265532209359514

Gaber, S., Cankar, G., Umek, L. M., & Tašner, V. (2012). The danger of inadequate conceptualisation in PISA for education policy. *Compare: A Journal of Comparative and International Education*, *42*(4), 647–663. https://doi.org/10.1080/03057925.2012.658275

Galaczi, E. D., ffrench, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach.

Assessment in Education: Principles, Policy & Practice, *18*(3), 217–237. https://doi.org/10.1080/0969594X.2011.574605

Glisan, E. W. (2012). National Standards: Research into practice. *Language Teaching*, *45*(4), 515–526. https://doi.org/10.1017/S0261444812000249

Green, A. (2018). Linking Tests of English for Academic Purposes to the CEFR: The Score User's Perspective. *Language Assessment Quarterly*, *15*(1), 59–74. https://doi.org/10.1080/15434303.2017.1350685

Green, R., & Wall, D. (2005). Language testing in the military: Problems, politics and progress. *Language Testing*, *22*(3), 379–398. https://doi.org/10.1191/0265532205lt314oa

Hamid, M. O., & Hoang, N. T. H. (2019). Humanising language testing. *TSL-EJ*, *22*(1).

Harvard University. (2019). *International Applicants*. Harvard College. https://college.harvard.edu/admissions/apply/international-applicants

Hatto, P. (2010a). *Standards and Standardisation. A practical guide for researchers*. European Commission.

Hatto, P. (2010b). *Standards and standardization handbook*. European Commission.

Hudson, T. (2012). Standards-based testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 479–495). Routledge.

Hulstijn, J. H. (2007). The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency. *The Modern Language Journal*, *91*(4), 663–667. https://doi.org/10.1111/j.1540-4781.2007.00627_5.x

Hyatt, D. (2013). Stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. *Journal of Further and Higher Education*, *37*(6), 844–863. https://doi.org/10.1080/0309877X.2012.684043

Hyatt, D., & Brooks, G. (2009). Investigating stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. *IELTS Research Reports*, *10*, 17–68.

IEA. (2019). *About IEA*. About IEA. https://www.iea.nl/

IELTS. (2019). *About IELTS USA*. IELTS. https://www.ielts.org/usa/about-ielts-usa

International Organization, & for Standardization. (2016). *How to write standards*. International Organization for Standardization.

International Organization for Standardization. (2018). *International standards & trade agreements*. World Standards Cooperation.

Jin, Y., Wu, Z., Alderson, C., & Song, W. (2017). Developing the China Standards of English: Challenges at macropolitical and micropolitical levels. *Language Testing in Asia*, *7*(1), 1. https://doi.org/10.1186/s40468-017-0032-5

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kaulfers, W. V. (1944). Wartime Development in Modern-Language Achievement Testing. *The Modern Language Journal*, *28*(2), 136–150. JSTOR. https://doi.org/10.2307/317331

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. https://doi.org/10.1111/j.1540-4781.2007.00627_6.x

Liskin-Gasparro, J. E. (1984). The ACTFL Proficiency Guidelines: Gateway to Testing and Curriculum. *Foreign Language Annals*, *17*(5), 475–489. http://dx.doi.org.kuleuven.ezproxy.kuleuven.be/10.1111/j.1944-9720.1984.tb01736.x

Liskin-Gasparro, J. E. (2003). The ACTFL Proficiency Guidelines and the Oral Proficiency Interview: A Brief History and Analysis of Their Survival. *Foreign Language Annals*, *36*(4), 483–490. https://doi.org/10.1111/j.1944-9720.2003.tb02137.x

Liss, J. M. (2013). Creative Destruction and Globalization: The Rise of Massive Standardized Education Platforms. *Globalizations*, *10*(4), 557–570. https://doi.org/10.1080/14747731.2013.806741

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy. *The Modern Language Journal*, *91*(4), 645–655. https://doi.org/10.1111/j.1540-4781.2007.00627_2.x

Little, D. (2019). Proficiency Guidelines and Frameworks. In J. Schwieter & A. Benati (Eds.), *The Cambridge Handbook of Language Learning* (pp. 550–574). Cambridge University Press.

Lowe, Jr., P. (1988). The unassimilated history. In P. Lowe, Jr. & C. W. Stansfield (Eds.), *Second language proficiency assessment: Current issues* (pp. 11–51). Prentice Hall Regents.

Massachusetts Institute of Technology. (2019). *IELTS Requirement for International Students*. MIT Media Lab. https://www.media.mit.edu/posts/ielts-requirement-for-international-students/

McNamara, T. (2014). 30 Years on—Evolution or Revolution? *Language Assessment Quarterly*, *11*(2), 226–232. https://doi.org/10.1080/15434303.2014.895830

Ministry of Education of the People's Republic of China. (2018). *National Language Standard. China's Standards of English Language Ability.* (GF 0018—2018). Ministry of Education of the People's Republic of China and National Language Commission of the People's Republic of China.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016. International Results in Reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement.

NATO. (2014). *Standardization Agreement STANAG 6001 Language Proficiency Levels*. Bureau for International Language Coordination.

Negishi, M., & Tono, Y. (2016). An update on the CEFR-J project and its impact on English language education in Japan. In C. Docherty & F. Barker (Eds.), *Language Assessment for Multilingualism. Proceedings of the ALTE Paris Conference, April 2014.* (pp. 113–134). Cambridge University Press.

North, B. (2014a). *English Profile Studies. The CEFR in Practice* (Vol. 4). Cambridge University Press. http://www.cambridge.org/gb/cambridgeenglish/catalog/teacher-training-development-and-research/cefr-in-practice

North, B. (2014b). Putting the Common European Framework of Reference to good use. *Language Teaching*, *47*(02), 228–249. https://doi.org/10.1017/S0261444811000206

North, B., & Piccardo, E. (2018). *Aligning the Canadian Language Benchmarks (CLB) to the Common European Framework of Reference (CEFR). Research report*. Centre for Canadian Language Benchmarks.

OECD. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. OECD Publishing.

OECD. (2019). *The Survey of Adult Skills: Reader's Companion, Third Edition*. OECD Publishing.

O'Loughlin, K. (2008). The use of IELTS for university selection in Australia. *IELTS Research Reports*, *8*(3), 145–241.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing*, *30*(3), 363–380. https://doi.org/10.1177/0265532213480336

Pearson. (2019a). *Score Comparison vs Other Tests for Researchers*. PTE Academic. https://pearsonpte.com/organizations/researchers/score-comparison-vs-competitors/

Pearson, W. S. (2019b). Critical perspectives on the IELTS test. *ELT Journal*, *73*(2), 197–206. https://doi.org/10.1093/elt/ccz006

Peirce, B. N., & Stewart, G. (1997). The development of the Canadian Language Benchmarks assessment. *TESL Canada Journal/ La Revue TESL Du Canada*, *14*(2), 17–31.

Princeton University. (2019). *International Students*. Princeton University Admission. https://admission.princeton.edu/how-apply/international-students

Randall, J. L., & Spolsky, B. (Eds.). (1975). *Testing Language Proficiency*. Center for Applied Linguistics.

Riazi, M. (2013). Concurrent and predictive validity of Pearson Test of English Academic (PTE Academic). *Papers in Language Testing and Assessment*, *2*(2), 1–27.

Ricardo-Osorio, J. G. (2008). A Study of Foreign Language Learning Outcomes Assessment in U.S. Undergraduate Education. *Foreign Language Annals*, *41*(4), 590–610. https://doi.org/10.1111/j.1944-9720.2008.tb03319.x

Rocca, L., Carlsen, C. H., & Deygers, B. (2019). *Linguistic Integration of adult migrants: Requirements and learning opportunities. Report on the 2018 Council of Europe and ALTE survey on language and knowledge of society policies for migrants*. Council of Europe.

Sarich, E. (2012). Accountability and External Testing Agencies. *Language Testing in Asia*, *2*(1), 26. https://doi.org/10.1186/2229-0443-2-1-26

Singer, J. D., Braun, H. I., & Chudowsky, N. (Eds.). (2018). *International education assessments. Cautions, conundrums, and common sense*. National Academy of Education.

Sjøberg, S. (2015). PISA and Global Educational Governance – A Critique of the Project, its Uses and Implications. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(1), 111–127. https://doi.org/10.12973/eurasia.2015.1310a

Sollenberger, H. E. (1978). Development and current use of the FSI Oral Interview Test. In J. L. D. Clark (Ed.), *Direct Testing of Speaking Proficiency: Theory and Application.* (pp. 1–13). Educational Testing Service.

Spolsky, B. (1995). *Measured Words: The Development of Objective Language Testing*. Oxford University Press.

Stanford University. (2019). *Exam Requirements for International Applicants*. Stanford Graduate Admissions. https://gradadmissions.stanford.edu/applying/starting-your-application/required-exams/exam-requirements-international-applicants

Stein, Z. (2016). *Social Justice and Educational Measurement.* Routledge.

Takala, S., Erickson, G., & Figueras, N. (2013). International Assessments. In *The Companion to Language Assessment*. John Wiley & Sons, Inc. http://onlinelibrary.wiley.com/doi/10.1002/9781118411360.wbcla052/abstract

Taylor, L. (2004). IELTS, Cambridge ESOL examinations and the Common European Framework. *Cambridge English: Research Notes*, *18*, 2–3.

The Guardian. (2014). OECD and Pisa tests are damaging education worldwide—Academics. *The Guardian*. https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics

The University of Cambridge, L. (2017). *English language requirements* [Text]. https://www.undergraduate.study.cam.ac.uk/international-students/english-language-requirements

The University of Chicago. (2019). *English Proficiency Testing*. College Admissions. http://collegeadmissions.uchicago.edu/apply/first-year-applicants/international-applicants/english-proficiency-testing

Trim, J. L. M. (2012). The Common European Framework of Reference for Languages and its background: A case study of cultural politics and educational influences. In M. Byram & L. Parmenter (Eds.), *The Common*

European Framework of Reference: The Globalisation of Language Education Policy (pp. 14–36). Multilingual Matters.

UNESCO. (2018). *Global Education Monitoring Report 2019: Migration, displacement and education – Building bridges, not walls.* UNESCO.

University College London. (2018). *English Language Entry Requirements*. International Students. https://www.ucl.ac.uk/prospective-students/international/applying-ucl/english-language-entry-requirements

University of California. (2019). *English language proficiency (TOEFL/IELTS)*. UC Admissions. https://admission.universityofcalifornia.edu/admission-requirements/international-applicants/english-language-proficiency-toefl-ielts.html

University of Oxford. (2019). *English language requirements*. University of Oxford. https://www.ox.ac.uk/admissions/undergraduate/applying-to-oxford/for-international-students/english-language-requirements?wssl=1

van Ek, J. A. (1975). *Systems Development in Adult Language Learning: TheThreshold Level in a European-Unit/Credit System for Modern Language Learning by Adults*. Council of Europe.

van Ek, J. A., & Trim, J. L. M. (1991a). *Threshold Level 1990*. Council of Europe.

van Ek, J. A., & Trim, J. L. M. (1991b). *Waystage 1990*. Council of Europe.

van Ek, J. A., & Trim, J. L. M. (2001). *Vantage*. Cambridge University Press.

Weigle, S. C., & Malone, M. M. (2016). Assessment of English for academic purposes. In K. Hyland & P. Shaw (Eds.), *The Routledge Handbook of*

English for Academic Purposes (pp. 165–177). Routledge. https://www.book2look.com/book/95iF77Y6oe

Weir, C. J. (2003). A survey of the history of the Certificate of Proficiency In English (CPE) in the twentieth century. In C. J. Weir & M. Milanovic (Eds.), *Continuity and innovation: The history of the CPE 1913-2002* (pp. 1–56). Cambridge University Press.

Woll, M. (1928). Standardization. *The ANNALS of the American Academy of Political and Social Science*, *137*(1), 47–48. https://doi.org/10.1177/000271622813700110

Wu, R. (2014). *Validating second language reading examinations: Establishing the validity of the GEPT through alignment with the Common European Framework of Reference* (Vol. 41). Cambridge University Press.

Yale College. (2019). *Applying to Yale as an International Student*. Yale College Undergraduate Admissions. https://admissions.yale.edu/applying-yale-international-student

Zhao, W., Wang, B., Coniam, D., & Xie, B. (2017). Calibrating the CEFR against the China Standards of English for College English vocabulary education in China. *Language Testing in Asia*, *7*(1), 5. https://doi.org/10.1186/s40468-017-0036-1

Zhao, Y. (2020). Two decades of havoc: A synthesis of criticism against PISA. *Journal of Educational Change*, *21*(2), 245–266. https://doi.org/10.1007/s10833-019-09367-x