Open access • Journal Article • DOI:10.1177/0265532217706196

University entrance language tests: A matter of justice: — Source link

Bart Deygers, Kris Van den Branden, Koen Van Gorp

Institutions: Katholieke Universiteit Leuven, Michigan State University

Published on: 01 Oct 2018 - Language Testing (SAGE PublicationsSage UK: London, England)

Topics: Language proficiency and Justice (ethics)

Article

LANGUAGE TESTING

Language Testing
1–28

DOI: 10.1177/0265532217706196
journals.sagepub.com/home/ltj
SAGE

University entrance language tests: A matter of justice

Bart Deygers
KU Leuven, Belgium

Kris Van den Branden
KU Leuven, Belgium

Koen Van Gorp
Michigan State University, USA

Abstract

University entrance language tests are often administered under the assumption that even if language proficiency does not determine academic success, a certain proficiency level is still required. Nevertheless, little research has focused on how well L2 students cope with the linguistic demands of their studies in the first months after passing an entrance test. Even fewer studies have taken a longitudinal perspective.

Set in Flanders, Belgium, this study examines the opinions and experiences of 24 university staff members and 31 international L2 students, of whom 20 were tracked longitudinally. Attention is also given to test/retest results, academic score sheets, and class recordings. To investigate the validity of inferences made on the basis of L2 students' scores, Kane's (2013) Interpretation/Use Argument approach is adopted, and principles from political philosophy are applied to investigate whether a policy that discriminates among students based on language test results can be considered just. It is concluded that the receptive language requirements of university studies exceed the expected B2 level and that the Flemish entrance tests include language tasks that are of little importance for first-year students. Furthermore, some of the students who failed the entrance test actually managed quite well in their studies – a result that entails broad implications concerning validation and justice even outside the study's localized setting.

Keywords

Justice, L2, language for academic purposes, mixed methods, university entrance test, validity

Corresponding author:
Bart Deygers, KU Leuven, Centrum voor Taal & Onderwijs, Blijde-Inkomststraat 7, Leuven, 3000, Belgium.
Email: bart.deygers@arts.kuleuven.be

Since its origins, centralized testing has been used to select individuals who possess skills that are deemed to be important for a future role or position (Spolsky, 1995). Bachman (1990) characterizes testing as an impartial way of distributing access to benefits or services, but for Foucault (1977), testing is an instrument of power that allows an in-group to select members from an out-group. Both Foucault and Messick (1989) have inspired language testers to examine critically the social consequences of tests and to question the gatekeeping functions they often perform (see also Shohamy, 2001). Recognizing the power imbalance involved, language testing organizations have developed principles to ensure that developers do not engage in activities that are inimical to candidates' best interests (e.g., ILTA, 2000). Set against this background, an overarching question of this study is "To what extent can language tests justifiably serve as gatekeepers to university entrance?"

University entrance language tests in Flanders

In Flanders, the Dutch-speaking northern part of Belgium, the default approach in first-year university classes is ex cathedra teaching, and students are generally not expected to speak or write much until the second or third years (De Wachter & Heeren, 2011). For Flemish L1 students, any diploma from a Flemish secondary school is a sufficient qualification for university entrance, and nobody is required to take a university entrance exam except prospective students of medicine and dentistry. Therefore, the actual selection process occurs at the end of the first year, when around 60% of students fail their exams (De Standaard, 2013). Failure is more likely for students with an atypical educational background (Smet, 2011), a low socio-economic status (De Wit et al., 2000) or an L1 different from Dutch (Lievens, 2013).

Unlike their peers with a Flemish secondary school diploma, prospective international L2 students are required to pass a language test, and two such tests – the ITNA and STRT – are accepted by all Flemish universities. In both tests, the oral component is administered face-to-face by a trained examiner, and, candidates are required to give a presentation based on slides, graphs and tables and provide an argument that supports or dismisses a prompt. However, the scoring rubrics differ, because the ITNA only takes into account formal linguistic criteria (e.g., vocabulary, grammar), whereas the STRT also focuses on content (i.e., whether a performance contains the main points mentioned in the prompt). The written component of the ITNA is computer-based and primarily includes selected-response question types, whereas the four paper-based writing tasks on the STRT require candidates to produce a summary or an argumentative text based on a listening or reading prompt. These STRT tasks are rated by two trained raters who score content and form, whereas the ITNA computer test is scored automatically using a binary rating scale. The ITNA does not include written production tasks, whereas the STRT does not contain non-integrated listening or reading tasks. Appendix 1 shows a detailed overview of the STRT and ITNA task types.

Both STRT and ITNA have been formally linked to the B2 level of the CEFR (Common European Framework of Reference for languages, Council of Europe, 2001) using the specification-standardization-validation approach described in Figueras, North, Takala, Van Avermaet, and Verhelst (2009). The CEFR describes six levels of

L2 proficiency, ranging from A1 (breakthrough) through C2 (mastery). The B2 level is commonly used as the threshold level for university entrance throughout Europe (Author, in press) and denotes an L2 user who can understand the main ideas of complex texts, interact fluently and spontaneously with native speakers, produce clear and detailed texts, and develop a sustained argument.

Validation and justice

Test scores convey little meaning in a contextual vacuum. A score only becomes *real* when it has real-life consequences, such as access to a valued position, service, or status. For that reason, most validation theories argue that validating a test without considering its social context and consequences is inadequate (Bachman & Palmer, 2010; Kane, 2013; Kane, Kane, & Clauser, 2017). What requires validation is not only the test itself (e.g., Borsboom & Markus, 2013), but also the way in which a score is interpreted and used (Kane, 2013). For Kane, validation is a matter of empirically investigating the claims that support the way in which score users interpret or use a score. Kane maintains that test validation falls primarily on score users and that test developers are accountable only for score use in contexts that they explicitly promote or could reasonably expect.

Kane has been criticized for assigning too much importance to score use rather than to the test itself (O'Sullivan, 2016), but he does not absolve test developers entirely. He specifically addresses the usefulness of content-based analysis to determine whether a test adequately samples from the target language use domain (Kane, 2013, p. 5). In the context of language for academic purposes (LAP), a substantial amount of primarily Anglo-American research has been devoted to identifying what typifies real-life academic language. There is general agreement that LAP requires advanced cognition and abstraction (Hulstijn, 2011; Taylor & Geranpayeh, 2011) and that argumentation, logic, analysis, and the ability to combine different sources and skills are central to the LAP construct (Cho & Bridgeman, 2012; Cumming, 2013). Furthermore, academic language involves specialized lexis (Snow, 2010; Hulstijn, 2011) and complex syntactical structures, including nominalizations, conditional structures, and embedded clauses (Gee, 2008; Snow, 2010; Hulstijn, 2011). Prototypical LAP tasks include giving presentations, describing graphs, understanding lectures, summarizing texts, and building an argument (Hyland & Hamp-Lyons, 2002; Lynch, 2011; Cho & Bridgeman, 2012). The threshold language level most associated with academic language proficiency in Europe is B2, but the debate remains whether a higher level might be more appropriate (Taylor & Geranpayeh, 2011; Hulstijn, 2011; Xi, Bridgeman, & Wendler, 2013).

This description of LAP may be too generic for a university entrance language test used within a specific context for a specific purpose (Lado, 1961), since local conventions may override general principles (Fløttum et al., 2006). Moreover, analyses of academic language proficiency in the Anglo-American tradition (Xi et al., 2013) will not necessarily apply to other contexts. Furthermore, LAP characterizes the language proficiency of an accomplished user of academic idiom, and not necessarily the language skills required of students embarking on their university studies. To identify real-life requirements, it is essential to determine what qualifies as a representative task in a specific context, and what does not.

IUA-based validation is the process of empirically determining whether real-world test score uses or inferences provide answers to specific problems within a specific context (Gorin, 2007; Kane, 2013). Kane demands strong empirical evidence and, when this evidence contradicts the claim made by the score user or the test developer, the validity argument cannot be maintained.

Universities who base entrance decisions on language test scores implicitly claim that students who pass can cope with the communicative demands of academic studies, and that those who fail are unlikely to be successful (McNamara & Ryan, 2011). The social consequences of this claim are profound, but are unsubstantiated in Flanders – as they are in many other contexts (McNamara & Ryan, 2011). Investigating them requires a perspective that is concerned with imbalances and inequities caused by testing policies. Theories of justice provide a valuable framework in this regard.

Post-Messick validity theorists like Kane have emphasized the importance of a test's social consequences, which has inspired discussions about fairness and justice (Davies, 2010; Kane, 2010; Kunnan, 2010; Xi, 2010; McNamara & Ryan, 2011). While there is general agreement that fairness primarily concerns bias and impartiality (McNamara & Ryan, 2011), justice has proven more elusive. Contrary to fairness, which presupposes the existence of a test, justice questions the legitimacy of using a test as a gatekeeper (McNamara & Ryan, 2011) since a test may introduce imbalance or inequity in a larger population (Kunnan, 2000). On this basis, the current paper examines a setting in which one subpopulation (international L2 students) is required to pass a test before being admitted to an institution that is open to others (students with a Flemish secondary school degree).

Much of what has been written about justice in language testing has been influenced by John Rawls (Davies, 2010). In Rawlsian political philosophy, fairness precedes justice, and the first principle of justice states that a ruling cannot be just if the foundation on which it is based is unfair, but fairness offers no guarantees for just rulings. The same applies in language testing: a test can be demonstrably fair while being indefensible as a policy instrument (McNamara & Ryan, 2011), whereas the opposite is hard to conceive. Rawls's second principle permits inequalities insofar as they work to the benefit of people who have an unfavorable starting position. Applying this principle to language testing is somewhat more challenging. Clapham (2000) calls for equal treatment in university admission testing by arguing that L2 university entrance tests should not include tasks that L1 speakers are not expected to perform. But, as can be deduced from Rawls's second principle, unequal treatment does not necessarily imply injustice (Dworkin, 2003). Universities may have sound reasons for demanding that L2 students possess linguistic competences that are not expected of their L1 colleagues. Consequently, a thorough context analysis will not necessarily yield a just testing policy. As a matter of fact, no preconditions can offer such guarantees, since justice might not be that absolute (Sen, 2010).

According to some, Rawls's presumption that true justice exists obstructs his theory's applicability. Sen (2010) therefore proposes an alternative in which justice is not seen as absolute, but as context-dependent: if a situation is perceived as unjust, and if freedom is restricted without a reasonable, rational argument, that situation *is* unjust. Dworkin (2003, 2011), a Rawlsian proponent of distributive justice, also supports the importance of freedom. For Dworkin, institutions are morally obliged to ensure equality of opportunity for

all their members – even when this implies unequal treatment. Rawls does not offer many practical guidelines for investigating justice, but his and Dworkin's work offer principles to evaluate the justice of a university entrance policy. Sen's reason-based approach blends with Kane's view of validation as hypothesis testing (Oller, 2012).

Based on the available definitions of justice in the language testing literature and on the insights drawn from Rawlsian political philosophy, this paper proposes, with Sen, to define justice as the absence of injustice. Hence, a policy that relies on tests for gatekeeping purposes can be considered unjust if it restricts test takers' freedom of access on grounds that are unreasonable or unsupported by empirical data.

The Flemish university entrance policy limits the freedom of access of L2 students based on the implicit claim that students cannot successfully participate in academic studies when they are below a certain language proficiency level. If this policy is just, people who fail the language test should not perform well in the TLU context. Otherwise, their freedom of opportunity would be unjustly limited, and the entrance policy would be indefensible. Irrespective of the differences between modern-day justice theories, it is unlikely that anyone would dispute the injustice of a policy that harms people in an already disadvantaged position, yet lacks rational or empirical grounds.

Research questions

Kane's approach to validation draws on Toulmin's (2003) argumentative model, in which every inference starts from evidence and results in a *claim*. Whoever makes that claim is required to provide justification (Kane, 2013, p. 12) by means of a *warrant* (i.e., a line of reasoning that connects the data to the claim). Warrants usually require support, or *backing*, and might need *qualifiers* to express the strength of the relationship. Since there may be conditions under which a warrant will not apply, Toulmin's model allows for *rebuttals*.

This study is based on three falsifiable claims and warrants, which are explicit formulations of implicit policy assumptions. The claims have been written so that university admission officers cannot reject them without dismissing the validity of their own policies. As required by Toulmin, the warrants are general statements connecting data to a claim. However, since Flemish university policy is not empirically based, this study relies on its own data.

The first claim is concerned with the sampling and operationalization of STRT and ITNA tasks.

Claim 1: *STRT and ITNA are representative of the real-life communicative demands of academic studies at Flemish universities.*

Warrant: If a test operationalizes the characteristics of academic language, and if it samples representative tasks, it can be used for university entrance purposes.

All Flemish universities require a B2 level and use ITNA and STRT as gatekeepers for L2 enrolment (KU Leuven, 2015; Universiteit Antwerpen, 2015; Universiteit Gent, 2015; Universiteit Hasselt, 2015). Thus, verifying whether the B2 level of these tests matches the university requirements is necessary in order to validate the second claim.

Claim 2: *Successful STRT and ITNA candidates are ready for the linguistic demands of academic studies at a Flemish university.*
Warrant: International L2 students who attain the B2 level can cope with the linguistic demands made on students entering university.

As the purpose of policy should be to advance social justice (Phillips, 2007), university entrance policy should not discriminate on grounds that are unsupported by empirical data or rational argumentation (Rawls, 2001; Sen, 2010). This principle lies at the heart of the third claim:

Claim 3: *Using ITNA and STRT as gatekeepers to university admission is just.*
Warrant: International L2 students who do not pass STRT or ITNA will not be successful since they cannot manage the linguistic demands made on students entering university.

To assess the justice of a university entrance policy, it must be determined whether freedom of access was duly restricted. In the context of this study, this means granting or denying access on the basis of an assessment of who stands a reasonable chance of success. Contrary to the first two claims, the third claim takes into consideration students' overall academic success

Method and procedure

This paper reports 24 faculty members' perceptions and 32 L2 students' experiences of the linguistic demands of university studies. With insights from needs analysis (Gilabert, 2005; Long, 2005) and mixed-method research (Creswell, 2015), sources, and methods are triangulated using a concurrent design that combines quantitative and qualitative data.

Participants

L2 students. This study is based on two groups of L2 participants at the three largest universities in Flanders representing the main research traditions (humanities, natural sciences, and social sciences). Classes ranged from 1,000-seat auditoria to smaller groups of about 50 students.

Group 1 ($L2_1$) Eleven L2 students attended their first semester at Ghent University during the academic year 2012–2013. They were enrolled in a non-obligatory course of L2 Dutch for academic purposes, which ended in December 2012. The interviews were conducted in a separate room during these classes.

The median participant age at the time of data collection was 20 years (range: 18–45), the median length of L2 Dutch instruction was 14 months (range: 9–48) and most participants (7) were female. Three of these participants had entered Ghent University at the bachelor's level and eight at the master's level. Since they were recruited after they had registered, they had already passed a language test (ITNA = 10, STRT = 1).

Participants will be referred to as S1–S11 individually, or collectively as $L2_1$ (Appendix 2).

Group 2 ($L2_2$) In the summer of 2014, 135 non-native speakers of Dutch planning to enroll at a Flemish university took both the ITNA and STRT as part of a concurrent validity study (Author, in press). Of this group, 68 candidates passed at least one of these tests, granting them university access. Less than half the group (32) registered for a Dutch-medium program, and 21 of these registrants agreed to participate. Before the start of the academic year, one student decided to postpone her studies for financial reasons, and so 20 participants remained. All 20 took both tests and seven of them received a different pass/fail outcome on each test. These students (except S28 – see below) were allowed to enroll despite having failed one entrance test. This is the first study to bypass the classic problem of truncated samples (Wall, Clapham, & Anderson, 1994) in this way. The problem is that students who do not pass an entrance test cannot enter university, so normally there is no way of knowing how they would have fared.

Ten participants were freshmen, and 10 were master's students. Six attended Ghent University, six the University of Leuven, and four the University of Antwerp. S15 attended an inter-university program. S28 failed the STRT and ITNA, but was able to register at the University of Hasselt, which accepts certificates from its own in-house B2 test. The median participant age at the time of data collection was 23 (range: 19–32), the median length of Dutch instruction was 11 months (range: 6–80), and the majority (17) were female. These participants will be referred to collectively as $L2_2$, or individually using their S12–S33 participant IDs. Appendix 3 provides additional information.

Data collection was carried out between October 2014 and July 2015. Attrition is typical in a longitudinal study, and seven students left the project by the end of the data collection. S26 and S27 dropped out in February 2015 to pursue studies in their L1. S25 quit in the same month because she had lost all motivation to pursue her studies. S28 (April 2015) and S29 (November 2015) had to give up because of visa issues. S30 and S31 left the project after one month without stating a reason.

University staff. In January and February 2014, 24 university staff members (out of 64 invited) each took part in one of six focus groups. The focus groups required information-rich participants (Reybold et al., 2013) who were able to provide knowledgeable insights (Patton, 2002) into the linguistic demands that are made on students at the start of university.

Purposeful participant selection (Freeman, 2000) was based on three inclusion criteria: affiliation, position, and experience. The participants represent the major universities (12 from Ghent University and 12 from KU Leuven) and the main academic traditions (humanities [6], natural sciences [7], and social sciences [7]). They represented both professors (15) (seven of whom were also directors of educational affairs) and tutors (6). Four participants worked at the central administration (language policy [2] and educational affairs [2]). When data were collected, the majority of the participants had substantial professional experience (experience at university: *Mdn*: 22 years, range: 3–35) and

teaching experience (participants not working in administration, teaching experience: *Mdn*: 19 years, range: 3–29; experience with first-year students: *Mdn*: 16 years, range: 3–29). These participants will be referred to as Ac1–Ac24 (see Appendix 4).

Although there were no direct professional ties between participants in the same focus group, hierarchic differences did exist, and power issues can make individuals change their views to match group consensus (Reybold, Lammert, & Stribling, 2013). For that reason, each focus group began with the collection of individual opinions in a paper-based questionnaire (Kahneman, 2011).

Data collection and analysis

L2 interviews and focus groups. All interviews and focus groups were conducted by the primary author, who was free to elaborate on salient subthemes that arose from a series of recurring must-ask questions. The data were audio recorded and transcribed in Dutch, but specific quotes were translated into English for this paper.

The interviews with the L2 students were conducted to determine whether these students felt ready for the linguistic demands of university (claim 2) and whether the academic language tasks they faced in real life matched the ones that were operationalized on the STRT and ITNA (claim 1).

The $L2_1$ interviews took place in October (the first weeks of the academic year) and December 2012, and dealt with the participants' experiences at university, the university's linguistic demands, the students' social network, and their perceived linguistic ability. $L2_2$ participants were interviewed during the academic year 2014–2015. Their perceptions of the linguistic demands at their university and the adequacy of their own language proficiency were a vital part of each interview, which focused on a different topic each time: the first weeks of university (October), classroom experiences (November), the first exams (February), the students' social network (March), and the participant's perceptions of the past year (June/July). The April interview was replaced by a retest of STRT.

The purpose of the focus groups was to come to a cross-disciplinary consensus (Belzile & Öberg, 2012) concerning the linguistic demands placed on students at the start of university, and to assess whether these demands matched the target level of the tests (claim 2). At the beginning, participants all agreed that the minimal linguistic demands were the same for all students, irrespective of their L1. They were then asked to estimate individually the relative importance of listening, reading, writing, and speaking skills. Next, they received three sets of four listening, reading, writing, or speaking samples (see Table 1), which they ranked in terms of difficulty. The agreed-upon order for every skill in every focus group corresponded with the CEFR levels assigned to the samples. Afterwards, as a group, they determined the minimal proficiency level they believed a first-year student should have, using an approach based on the bookmark method, a frequently used standard-setting procedure (Bérešová, Breton, Noijons, & Szabó, 2011). Table 1 presents an overview of the samples used in the focus groups (excluding the speaking samples, as they will not be referred to specifically in this article) and identifies the source, the topic, the length, the percentage of low-frequency ($\geqslant$ 5000) and high-frequency words ($\leqslant$ 2000), the difficulty (as measured by Flesch-Douma (FD), and the

Table 1. Focus group samples, arranged by CEFR level.

		Code	Source	Topic	Length	⩽2000	⩾5000	FD	W/m
Writing	B1	W3	L2 test performance	Law	72	88.2%	8.7%	59	
	B2	W1	L2 test performance	Advertising	186	83.2%	7.3%	52	
	C1	W2	First-year paper, L1	Arabic studies	170	78%	4.7%	61	
	C2	W4	Dissertation, L1	Engineering	121	75.4%	13.2%	40	
Reading	B1	R4	B1 test	History	163	74.4%	7.3%	81	
	B2	R1	B2 test	Musicology	177	79.1%	13%	55	
	C1	R2	C1 test	Linguistics	179	79.8%	11.3%	29	
	C2	R3	Course book	Sociology	159	70.8%	21.1%	4	
Listening	B2	Li4	B2 test	Biology	2.14	76.3%	8.8%		147
	C1	Li2	C1 test	Physics	1.56	84.6%	10.3%		126
	C2	Li1	Radio lecture	Mathematics	2.03	86%	8.2%		145
	C2	Li3	University lecture	Philosophy	2.01	82.9%	10.3%		116

Length: in words (writing and reading) or minutes (listening).
⩽2000: high-frequency words.
⩽5000: low-frequency words.
FD: Flesch-Douma readability: 100 is very easy; 0 is very difficult.
W/m: words per minute.

CEFR level of the samples. Word frequencies, readability indices, and speech rate were used as indicators of complexity to supplement the CEFR level assigned to the samples. All L2 samples (W1, W3, R1, R2, R4, Li2, Li4) were selected from a sample bank containing L2 performances and tasks that were linked to the CEFR by an independent committee of experts (Nederlandse Taalunie, 2015) following the procedures outlined in Figueras et al. (2009). The L1 writing samples were chosen by academic writing tutors, who were asked to provide a representative performance from a first-year (W2) and final-year (W4) student. The authentic reading sample (R3) from the first chapter of a first-year sociology textbook was considered representative by the participants. Sample Li1 was selected from a radio broadcast in which a professor explains a mathematical problem to a wide audience of non-specialists, while Li3 was recorded purposefully with a philosophy professor, who was asked to teach his introductory class. Samples W2, W4, R3, Li1, and Li3 were linked to the CEFR by four experienced members of the above-mentioned committee.

All transcriptions were coded a priori and inductively (Dey, 1993; Miles & Huberman, 1994) using NVivo 11. The a priori coding schemes were based on salient themes that emerged from the LAP literature review, on the interview and focus group scenarios, and on research into L2 students' experiences at Flemish universities (De Bruyn, 2011). During coding, themes emerged that were not foreseen in the a priori scheme, adding an inductive layer of analysis (Glaser & Strauss, 1967). In order to check the coding consistency, a research assistant recoded one focus group and all $L2_2$ interviews conducted in November 2014, using the a priori coding scheme. The inter-coder agreement was substantial (Landis & Koch, 1977): L2 interviews $K_w = .62$, L1 focus groups $K_w = .60$, or an exact agreement of >90%.

Academic language skill questionnaire. The university staff participants were asked to fill out a questionnaire in which they selected the most important academic language skills for first-year students. As the perceptions of academics may differ from students, the $L2_2$ informants received the same questionnaire in February 2015. The views of the academic participants and the opinions of the $L2_2$ informants were used to assess whether the task selection in STRT and ITNA was representative of the actual linguistic demands at Flemish universities (claim 1).

The list of language skills used in the questionnaire was based on prior research and commonly occurring task types in 13 European tests that grant access to higher education (CELI 3, CELI 4, Studieprøven, Test i norsk – høyere nivå, Staatexamen NT2 II, ITNA, PTHO, PAT, IELTS, DALF, TCF, TELC C1 Hochschule, TestDAF). Skills featured in at least seven of these tests were included. Participants were free to add to the list, which happened on one occasion ("accurate expression of ideas"). The categories in the list (see Table 2) were purposefully broad, because they had to be meaningful to non-linguists (Long, 2005).

Complementary data sources. Long (2005) and Gilabert (2005) recommend supplementing interview and focus group data with other sources in order to get a complete picture of the phenomena under examination. For this study, the following complementary data were collected:

Class recordings and field notes

In November 2014, the researcher attended a class of each $L2_2$ student's choosing. Eleven lecturers gave permission to have their class audio recorded. The researcher also took field notes before, during, and after the classes.

These data were used to compare test tasks to real-life language situations (claim 1). Additionally, the first 30 minutes of each class were transcribed and analyzed for speed (words/minute) and word frequency (using *TST Centrale*, a lemma-based corpus for Dutch). They were then compared to STRT audio prompts, allowing for a comparison between the linguistic demands of university lectures and the STRT audio prompts (claim 2). Lastly, the field notes were analyzed for instances that showed whether a participant was able to cope linguistically during class (claim 2).

Academic score transcripts and test/retest scores

In April 2015, the remaining $L2_2$ participants ($N = 15$) took two STRT test tasks again: writing a summary of a scripted lecture about industrialization and giving a 10-minute presentation about pollution, based on slides. As it was impractical to administer the whole test again, the two tasks that explained most of the overall score variance in the previous test administration ($N = 913$) were selected for the retest ($R^2_{adj} = .91$, $p < .000$; summary $\beta = .52$, $p < .000$ presentation $\beta = .57$, $p < .000$). If the $L2_2$ students' language ability were to improve between the test and the retest, it could be argued that even if the test level does not represent actual academic demands, L2 students do make progress linguistically in the course of the year. If true, this could be used as a qualifier in the warrant of claim 2.

Table 2. Academic language skills selected in focus groups ($N = 6$).

Academic language skills	#	+	2+	3+	4+	5+
Express ideas accurately	6	1	1	1	2	1
Understand coherence & cohesion	5			1	1	3
Take class notes	5		1	1	2	1
Compose a logical argumentation	3	1		1	1	
Grammatical accuracy	3	1	2			
Summarize long text	2		1	1		
Master academic vocabulary	1					1
Understand scientific text in detail	1			1		
Understand scientific text as a whole	1	1				
Look up information	1	1				
Describe graphs & tables	0					
Summarize multiple sources	0					
Understand implicit message	0					
Give a presentation	0					

# indicates times selected.
+ indicates times awarded level of importance (5+ is most important).

At the end of the second semester (July 2015), the $L2_2$ participants provided the researcher with transcripts of their academic results. Based on their academic success, the participants were divided into two groups: students who had passed at least half of their courses ($L2_2^+$, $N = 8$) and those who had not ($L2_2^-$, $N = 8$). The $L2_2^-$ group did not include the two students who left university because of immigration problems or the two students who left the project early. The academic performance data were combined with the entrance test results to assess whether any academically successful L2 students had failed the STRT or ITNA, as failure would have kept them from enrolling (claim 3).

Given the small number of participants and the non-normal distribution, non-parametric tests were used to analyze these data. Wilcoxon's Signed Rank Test and effect sizes were used to determine whether $L2_2^+$ students had achieved higher initial STRT or ITNA or ITNA scores, and to measure score gains on STRT tasks. As the tests' CEFR-based scales may be too broad to measure gains over a matter of months, more detailed analyses were conducted based on a methodological approach adopted by Serrano, Tragant, and Llanes (2012) and Llanes, Tragant, and Serrano (2012). This analysis relies on comparing measurements of complexity (lexical: type/token ratio; syntactic: clauses/T-unit), accuracy (written: errors/T-unit; oral: errors/AS-unit), and fluency (written: words/T-unit; oral: pruned syllables/minute) over time. These results will be referenced below, but the analyses themselves have been reported in detail elsewhere (Author, in press). All quantitative analyses were conducted with R Studio (*QuantPsyc* and *car* packages).

Results

Claim 1: STRT and ITNA are representative of the real-life communicative demands of academic studies at Flemish universities

The data used to verify this claim are as follows:

- the $L2_2$ participants' opinion of the tests' representativeness;
- the results of the skill ranking exercise in the focus groups;
- the experiences of $L2_1$ and $L2_2$ participants; and
- the results of the academic language skill questionnaire in the focus groups and in the $L2_2$ interviews.

In October 2014, when asked which test they preferred, six $L2_2$ participants chose ITNA, 10 chose STRT, and five were undecided. Participants who preferred STRT often did so because they felt that ITNA's computer component lacked content representativeness: four students disliked ITNA's selected-response tasks and six disapproved of the absence of writing tasks. According to seven participants, ITNA's vocabulary tasks were not representative for real-world university lexis.

The $L2_2$ participants perceived the ITNA and STRT listening tasks as the most useful, albeit not entirely representative. The importance of listening is reflected in the academic participants' skill ranking results. There was overall consensus that for first-year students, receptive skills are more essential than productive skills, and that speaking is of little importance.

> Ac4: Speaking just does not happen in the first year … First and foremost, students entering university should be able to store information.

All $L2_2$ participants also judged receptive skills to be the most important:

> S1: I mainly have to listen basically … I actually have the feeling that my Dutch is getting worse. For my courses I don't need to write much. I mainly write down formulas, but that doesn't require much language, so I don't practice anymore.
>
> (December 2012)

After two months at university, four $L2_2$ participants reported speaking Dutch quite often. Others had rarely used it (5), were afraid to speak it (5), or had not used it yet (4). Ten of the 11 $L2_1$ participants claimed they "hardly ever" spoke Dutch at university. A few students in this study were involved in group work, which typically involves speaking, yet some found ways to avoid speaking by using chat (S11) or email (S15).

> S15 I do everything I can to prevent a meeting with students … I always write long texts to give my opinion, but in a meeting all I can say is yes, no and OK.
>
> (November 2014)

S15 hints at the importance of speaking in gaining acceptance in a community of peers and building an identity in a new context (Morita, 2004; Amuzie & Winke, 2009). Identity and acceptance were major recurring themes in the $L2_2$ interviews, but they are beyond the scope of this paper.

Having established the relative importance of receptive and productive skills, the university staff participants took the questionnaire to decide which academic language skills were most important for first-year students. Table 2 indicates the relative importance assigned to each skill in the six focus groups.

The consensus in every focus group was that using meaningful language is the most important language skill for first-year students.

> Ac20 If the message is correct, it's ok … What I understand as "meaningful" is very basic language: I have to be able to agree or disagree with what is being said.

For the university staff, the second most important academic language skill was "*understand coherence and cohesion*," which was defined as being able to distinguish essential from non-essential information (Ac4, Ac6, Ac8, Ac17), receptively, but also productively. Even though the university staff considered receptive skills to be essential, their selection also included skills for passing written examinations, such as writing down answers in an accurate and structured way.

When the $L2_2$ participants received the questionnaire in February 2014, their selection reaffirmed the importance of receptive skills such as "*understand academic lexis*," "*understand implicit message*," and "*understand scientific text as a whole*." "*Compose a logical argumentation*" and "*take class notes*" occurred in the top five of both groups.

In five focus groups, the consensus was that students can start university studies without having acquired specific academic lexis because introducing it is the lecturer's task. Yet every $L2_2$ informant complained that limited lexical knowledge was a major obstacle. In most cases, $L2_2$ participants were not referring to highly specialized terms, but to words that are commonly acquired in the course of Flemish secondary education. It is possible that the university staff underestimated the lexical complexity of their own language use, in assuming that all students would know frequently used words within their field. It is clear from the excerpt below that this assumption may be misguided. Like other L2 participants involved in this study, S13 was unfamiliar with basic mathematical terminology at the start of university.

> S13 Belgian students know these words from high school, from basic maths or something – it's not that hard. But when your vocabulary is not adjusted, you need to think "infinite, what is infinite?" And you need to think in numbers, and when I think in numbers, I think in Spanish.
>
> (October 2014)

The skill "*understand implicit messages*" was also perceived differently by academic staff and $L2_2$ students. Professors were convinced that "academic language is not supposed to be implicit" (Ac 7), but for $L2_2$ students, implicit language includes irony, jokes, and idioms – all of which are important when attending lectures. During these lectures, most $L2_2$ participants took notes, a skill considered important by L2 students and

university staff. But – as both groups acknowledge – note-taking does not mean writing full pencil-and-paper summaries as operationalized in STRT. More than two-thirds of the $L2_2$ participants wrote "comments on a hand-out" (Ac22) without taking actual notes.

The skills the participants did not select are at least as important as the ones they did. All $L2_2$ participants and all university staff members disregarded the skills "*give a presentation*" and "*describe graphs and tables.*" Nevertheless, delivering a presentation is one of the two tasks included in the oral components of the STRT and ITNA, and at least two STRT tasks rely on candidates being able to describe graphic or tabular input.

The following section focuses not so much on the content of the tests, but on the connection between the required language proficiency level and the real-world expectations.

Claim 2: Successful STRT and ITNA candidates are ready for the linguistic demands of academic studies at a Flemish university

The data used to assess the second claim are as follows:

- test/retest scores, in order to measure differences in L2 proficiency over time;
- focus group discussion data about the listening, reading and writing samples (Table 1), in order to determine the minimal level of competence that the academic staff members expected (though speaking samples were not part of the discussions, as all groups agreed that it was the least important skill for first-year students to master);
- interviews with L2 participants, in order to cross-check the focus group results and to provide concrete examples of the linguistic hurdles they faced;
- field notes, in order to provide first-hand observations of how L2 participants experienced lectures; and
- a comparison of the lexical demands and speed of actual lectures and STRT listening tasks in order to determine whether the participants' perceptions were confirmed by actual observations.

Listening. The focus group participants ranked the samples (Table 1) before determining the minimally expected level. In all focus groups, it was decided that samples Li1 (C2) and Li3 (C2) were the most demanding, but also the most representative because they contained an argumentative component and because they were live recordings of lectures delivered in a natural way. The focus groups further agreed that Li3 is above what can be expected from a student on day one because it relies on prior content knowledge. Li1 was considered lexically less demanding, but with a high information density and a straightforward line of reasoning. The groups decided to put the cut-off point between Li1 and Li3. The B2 sample (Li4) was labeled as idealized, unrealistic, and unrepresentative because of its straightforward structure, its monothematic nature and its "cleanness."

> Ac8 No professor teaches like sample 4. It's too clean …
>
> Ac5 I agree. It was secondary school talk.
>
> Ac6 Like a television program for primary school children.

Confirming the university staff's intuition, all $L2_1$ participants struggled to understand the natural, unpolished language of university lectures.

> S3 The professor speaks too fluently for me and too academic … I try to understand but it still is hard. I am always in doubt. What did he say? What did he say, I always wonder.
>
> (October 2012)

Some L2 participants dropped out (S2), quit going to classes (S26, S27), or experienced loss of motivation (S16, S21, S24) primarily because they had problems with understanding lectures. Most $L2_2$ participants felt unprepared for the listening demands of university lectures. Of the four participants who reported no listening problems, three gave up before the end of the year. The main obstacles to understanding lectures had to do with pronunciation, intonation, and pace (11), regional accents (9), and jargon, idioms, and jokes (9).

> S26 [The professor] has the worst accent, so I don't understand anything. Nothing. Thank goodness we have a syllabus.
>
> I Does it have to do with the content of the course is it the language?
>
> S26 I don't know, do I? I just bought the syllabus and I will discover what it is about.
>
> I So you really don't understand anything?
>
> S26 Seriously. Nothing.
>
> (October 2014)

At the end of the year, seven $L2_2$ participants felt quite sure that they understood classes better than at the start of the year, although unfamiliar accents or unclear pronunciation remained a problem for most.

> S21 During the first semester it was not easy to understand a professor, but the second semester is better. I can understand well now. Not everything, but the most things. I can understand other students, but not people who do not articulate well.
>
> (June 2014)

The interviews showed that lexical problems caused additional difficulties during lectures. A comparison (see Table 3) between the language used in eight scripted lectures used as STRT prompts and in 12 actual university lectures confirmed this: lectures contained more low-frequency words than the prompts. Contrary to the perception of the L2 students, however, the average pace of real-life lectures was slower than the test prompts.

The field notes reveal other, more qualitative differences between the test prompts and in-class experiences. All bachelor's and master's-level classes the researcher attended in the course of this study, whether they were attended by 50 or by 500 students, were primarily ex cathedra. In some classes, professors asked an occasional question, but there was never any sustained interaction. In most classes, there was a lot of background noise:

Table 3. University lectures and STRT listening prompts.

		1K–2K[a]	5K–7K	7K+	W/m[b]
Test (*N* = 8)	*M*	5.93	2.33	6.50	148.33
	SD	4.43	2.52	2.78	18.04
Class (*N* = 12)	*M*	6.67	1.23	10.37	103.86
	SD	3.79	1.08	5.82	18.60

[a]% of words used in frequency band.
[b]Words per minute.

"there is a constant buzz of students talking among each other during class. The professor just talks through the noise" (Field notes S13, p. 3). In one class, the distractions were particularly intrusive: "students around us are drinking bourbon, there's a lot of talking, screaming, and shouting" (Field notes S20, p. 1).

Reading. The academic participants unanimously considered reading sample R3 (C2) the most demanding. In all focus groups, individual members suggested putting the cut-off score above R3 because it represented the actual language of syllabi. In the end, it was agreed that it was unrealistic to expect students entering university to cope with texts of this level, although they would encounter such texts early on in their studies. The groups also agreed that prospective students should have mastered R2 (C1), but not the structurally complex R3.

In every focus group, texts R1 (B2) and R4 (B1) were considered below the mark. Participants claimed that "students who can only master text 1 have a problem" (Ac13) and that text R4 is "annoyingly transparent" (Ac7). These texts were considered too easy because of their clear structure, low information density, and comparatively simple development of ideas.

For the $L2_2$ participants, reading presented a problem, but one they mostly managed. All $L2_2$ participants reported that reading took much longer in Dutch than in their L1 because they looked up words, consulted sources in English or in their L1 to understand concepts they did not grasp in Dutch, or translated parts of their courses. At least three participants had translated their entire courses into their L1.

> S28 In all honesty, I'm a bit of a maximalist … I lose a lot of time by translating.
>
> I Do you translate your courses?
>
> S28 Nearly everything yes: some of the words overlap. But the rest is different. I can't study in Dutch, but in Armenian I just need to read it once or twice and I know it.
>
> (February 2015)

As the year progressed, quite a few $L2_2$ participants reported an improvement in reading comprehension (S22, S23, S26, S27) or speed (S13, S14, S28). Other participants (S16, S24) confirmed that their reading had improved, but was not up to standard yet.

S16 There is one book about stuff Freud wrote – very difficult language … I try to read it, but do not understand it.

Writing. The focus groups put the cut-off point for writing above W1 (B2) and below W2 (C1). Poor text structure and syntax were the reasons why the final cut-off point was set above W1, even though university students do hand in texts at this level: "[W1] is representative of what many students do" (Ac17). In some groups, even W3 (B1) was not considered uncommon, nor was written language at this level seen as a reason to fail a student – even though it was substandard.

Ac21 If you ask me whether this person may enter university, I'd say no. If you ask me whether somebody could pass my course if he or she writes like this: well, yes. If he or she writes factually correct answers I'd feel obliged to pass this person.

In line with the views expressed in the focus groups, the $L2_2$ participants generally found writing difficult and time consuming, but not necessarily problematic. Many students developed effective coping strategies, such as asking for permission to write exams in English. Students who were involved in group work found that L1 students often corrected their texts. Others had not yet received a writing assignment and had only taken multiple-choice exams. Quite a few $L2_2$ participants did not assume that their written skills had improved since the start of classes. Some even felt that their written Dutch had become worse (S25, February 2014).

Test/retest scores. The academic participants expected that L2 students who passed the tests would not necessarily possess the required proficiency level. Nevertheless, they assumed that L2 students' language proficiency would improve as the year progressed. Contrary to these expectations, however, the STRT retest yielded only negligible effect sizes and non-significant gains, as measured by the tests' CEFR-based rating scale, both for the whole group (Writing: $W = 31$, $p = .159$, $r = -.314$; Speaking: $W = 43.5$, $p = .824$, $r = -.052$) and for the academically successful subpopulation (Writing: $W = 11.5$, $p = .331$, $r = -.280$; Speaking: $W = 16.5$, $p = .872$, $r = -.046$). More detailed analyses of the performances (Deygers,unpublished data) indicated that there were no significant gains on either task in terms of lexical or syntactic complexity, accuracy or fluency, with small effect sizes r (−.01 – .17). As STRT is an integrated-skills test, it does not directly measure listening and reading, but when a salient point from the prompt is mentioned correctly in the candidate performance, one point is awarded. To the extent that STRT's integrated tasks measure receptive skills, no significant progress was recorded (written $W = 37$, $p = .206$, $r = -0.28$; oral $W = 48.5$, $p = .505$, $r = -.156$).

Claim 3: Using ITNA and STRT as gatekeepers to university admission is just

As university entrance largely depends on STRT or ITNA certificates, students who fail either of these tests would be unprepared to participate successfully in academic studies.

This study included candidates who entered university after failing one test but passing the other, which provided the opportunity to assess the validity of claim 3. The following sources of data were used to assess this claim:

- the participants' perceptions about the justice of the university entrance policy;
- the initial STRT and ITNA outcomes, presented in Appendix 3; and
- indicators of academic success (i.e., $L2_2^+$ and $L2_2^-$) for participants who had attained more or less than 50% of the credits in their program.

Most participants agreed that the use of a language test as a gatekeeper to university entrance was warranted. The consensus among university staff was that low linguistic entrance requirements create false expectations. They felt that L2 entrance requirements needed to be high because there are virtually no support systems for L2 students (Ac24), and because they "are in the auditoria with other students [and] it's better to give these students a clear message from the start" (Ac17). Most $L2_2$ participants also supported the use of a university entrance language test, but contrary to the university staff, they did not feel the need to raise the required entrance level, because it would deny too many L2 students the chance to start. Only one $L2_2$ participant opposed L2 university entrance tests: "Somebody can find the language easy, but be super stupid academically. He won't succeed, but the opposite can also be the case" (S27).

The academic results of the $L2_2$ participants seem to partially confirm S27's point: there is no clear link between language test scores and academic success. $L2_2^+$ students did not significantly outperform $L2_2^-$ students on the initial STRT and ITNA tests (STRT: $W = 46$, $p = .625$, $r = -.115$; ITNA: $W = 51$, $p = .599$, $r = -.120$). Likewise on the STRT retest, $L2_2^+$ did not outperform $L2_2^-$ (STRT writing: $W = 14$, $p = .741$, $r = -.104$; STRT speaking: $W = 16$, $p = .451$, $r = -.238$). When interpreting these outcomes, it is important to note that only 13 $L2_2$ participants took part in the final exams of that year, so considering the actual impact of the policy in absolute numbers may provide a clearer picture.

Participants who were academically unsuccessful, yet gained admission on the basis of a language test, can be considered as false positives, in the sense that they gained entrance to university, but were, for various reasons, unable to complete successfully the first year; whereas participants who failed STRT or ITNA, yet belonged to the $L2_2^+$ group, can be considered as false negatives. Of the 16 participants, the STRT and ITNA respectively assigned seven and six false positives. However, as false positives do not lead to the exclusion of members of a specific group, they do not qualify as an injustice. From a justice perspective, false negatives carry considerably more weight. In the $L2_2^+$ group, ITNA assigned two (S15, S16) false negatives, STRT none. S15 was not a confident speaker, but passed her exams with honors. S16 had experienced a difficult first semester, but passed all of the second semester exams. It is interesting to note that S28 had failed both STRT and ITNA and had then registered at a smaller university after passing their in-house test. She managed quite well at this university, but did not finish the year because of immigration issues. But before S28's study visa was repealed and she was extradited, she had a better-than-average academic score sheet and had outscored quite a few L1 colleagues on *Business Dutch*. If S28 is included in the $L2_2^+$ group, the sum of false negatives is three for ITNA and one for STRT. Had the new STRT cut score been in effect at the time of data collection, that test would have assigned two false negatives.

Discussion

Claim 1: STRT and ITNA are representative of the real-life communicative demands of academic studies at Flemish universities

To some extent, the STRT and ITNA are representative of the communicative demands of academic programs at Flemish universities. STRT takes into account content-related criteria, which corresponds to the importance that the university staff assigned to meaningful rather than correct language. However, ITNA only considers linguistic correctness. Both tests take into account the importance of lexis; in STRT and in the oral component of ITNA the use of appropriate vocabulary is a rating criterion. ITNA also tests vocabulary knowledge in selected-response tasks, but this task was most often identified as the least representative by the $L2_2$ participants. The $L2_2$ participants and the university staff agreed on the importance of logical argumentation and taking class notes. Both skills are operationalized in STRT, but the way note-taking is conceptualized does not entirely take into account Power Point-based teaching, which is referred to by the participants and supported by research (e.g., Lynch, 2011).

In some cases, the operationalization of the STRT and ITNA contrasts with real-life demands. The university staff participants and the L2 students at the bachelor's and master's level agree that for students at Flemish universities receptive skills are more important than productive skills. Oral skills are considered least important. It is striking that all $L2_2$ participants and university staff members considered giving a presentation and describing graphs and tables unimportant skills. This does not necessarily mean that productive skills should not be assessed in university entrance tests, as oral skills matter for students' social integration (Morita, 2004; Amuzie & Winke, 2009), and written skills are important for passing examinations. What these observations do imply is that productive skills are generally less important than receptive skills for Flemish university students, especially in their first year. Consequently, assigning decisive importance to oral proficiency tests (ITNA) or relying on productive output alone (STRT) might not correspond to real-life demands.

The test developers' approach to academic language does not align with the linguistic reality at Flemish universities. It appears that the test developers have drawn largely on the LAP literature, which is primarily Anglo-Saxon, without taking into account the specific features of the Flemish context. Consequently, a more suitable warrant to the first claim could be formulated as follows: "*If a test adequately simulates the characteristics of academic language within the target context, and if it adequately operationalizes representative tasks, it may be valid for university entrance purposes.*"

Claim 2: Successful STRT and ITNA candidates are ready for the linguistic demands of academic studies at a Flemish university

The university staff participants agreed that L2 students would inevitably be less proficient than their L1 colleagues at the onset of their studies, but a commonly held assumption was that by attending classes, L2 students would become more proficient at Dutch. This study did not generate any empirical evidence to support this hypothesis. The STRT retest in April 2015 did not yield any significant score gains, or gains in terms of

complexity, accuracy, or fluency (for similar finding, see Kinginger, 2008; Amuzie & Winke, 2009; Dewey et al., 2014). The assumption that L2 students' language proficiency will increase over a semester simply by attending classes in Dutch thus seems unlikely. Consequently, it could be argued that it is vital for L2 students entering university to have achieved a language proficiency level that matches the linguistic demands of the TLU context. The results show that this is not the case, especially for listening.

Instead, this study shows that the real-life demands regarding listening and reading skills are considerably higher than those for writing or speaking. The university staff and the $L2_2$ participants referred to the B2 STRT listening prompt as an unrealistic idealization. The scripted lecture used in STRT did not contain the regional variations, information density, structural flaws, idiosyncratic accents, or disruptions that make it hard for L2 students to understand authentic university lectures. Therefore, few L2 participants felt prepared for academic listening demands. With one or two exceptions, all L2 participants experienced problems understanding academic lectures. This outcome confirms previous research, which found that B2 listeners are able to understand far less of an academic lecture than is usually assumed (Field, 2011; Lynch, 2011).

The fact that most participants reported listening as the most problematic skill does not imply that they were adequately proficient in the other skills. Listening simply posed the most immediate threat, and their repertoire of coping strategies was fairly limited. The university staff participants also considered the B2 reading samples unrepresentative, and all L2 participants reported problems with reading. For many students, this implied that they had to study twice as long as they did in their L1, or had to translate coursework to their L1. $L2_2$ participants also reported problems with writing, but often experienced some leniency from professors or assistance from L1 peers. Given their reported struggles, it can be somewhat surprising that the L2 students preferred not to raise the level of the entrance test. For them, however, raising the level implied giving fewer international students the chance to register for university, which is relevant to the justice of the admissions process (see below).

This study offers very little – if any – data to back the claim that students who pass the language test are able to cope with real-life linguistic demands. All L2 students included in this study had passed the ITNA or STRT or both (except for S28 – see above). Some managed remarkably well, but the majority of L2 participants were not ready to deal with the linguistic demands of academic studies at university (see Römhild et al., 2011). Additionally, this study affirms Hulstijn's (2014) assertion that in academic contexts, uneven language proficiency profiles are the rule. The data do not suggest that a B2 requirement for every skill corresponds with the actual language requirements at Flemish universities, and as such neither the warrant nor the claim withstands close scrutiny.

Claim 3: Using ITNA and STRT as gatekeepers to university admission is just

Carlsen (in press) distinguishes two kinds of interpretations given to university entrance language test scores. The *strong* interpretation implies that students who pass a test are ready for the linguistic demands of university. This study shows that students with high language test scores were not guaranteed to be successful. As there is little if any research to suggest otherwise (e.g. Lee & Greene, 2007; Cho & Bridgeman, 2012), the strong

interpretation was not a hypothesis this study was designed to test. The *weak* interpretation however is at the basis of many university entrance policies, including the Flemish one. It assumes that students who do not pass a language test are not ready for the linguistic demands of university, and will therefore be unlikely to achieve academic success. This interpretation, which is based on the idea of a minimally competent user, serves as the warrant to the third claim in this study.

The idea that students who do not meet the minimum language requirements will not manage in real life offers the rationale for restricting L2 students' freedom of access. Investigating this is difficult, however, because it is often impossible to trace false negatives. In the design of this study, the problem of truncated samples (Wall et al., 1994) was bypassed by tracking seven $L2_2$ participants who had actually failed the STRT or ITNA. Out of these seven people, ITNA assigned three false negatives, STRT one (two, using the new cut score). These absolute numbers may seem rather small, but when it concerns high-stakes claims, Kane does not allow for any negative evidence. In the context of admission testing, false negatives signal an unfounded restriction of access that applies to one subpopulation alone. According to leading justice theorists, an empirically unsupported restriction of opportunities is cause for concern (Rawls, 1971, 2001; Dworkin, 2003, 2011; Sen, 2010). Consequently, in the case at hand the presumption of a just policy cannot be upheld.

The study also found a substantial proportion of false positives for both tests, which – while not qualifying as an injustice – are not necessarily unproblematic. When students are admitted with language skills below the real-world expectations, it does not benefit the student or the university. It would seem that a university has a responsibility towards these students, which they could meet by helping L2 students reach the real-world expectations through needs-based curricular language courses (see Byrnes, Maxim, & Norris, 2010).

Conclusion and implications

The results of this study reveal that the content of the Flemish university entrance tests at times deviates strongly from real-life language demands, that students who passed these B2 tests were not ready for the receptive linguistic demands of academic studies at university, and that one in five participants would have been unjustly denied university entrance on the basis of their ITNA result or the most recent STRT cut score. Consequently, when considered as a whole, the data presented in this study do not validate the claims on which the university entrance policy rests. More generally, the results underscore the importance of a thorough TLU analysis, question the position of the B2 level as the minimum overall requirement for university entrance tests, and highlight the importance of a focus on justice in high-stakes testing.

Hume's (1978) insistence that we are under a moral obligation to act once an injustice is known suggests implications to this study. One is local: the current L2 university entrance policy in Flanders should be re-examined with regard to validity and justice. Implementing different CEFR-requirements for receptive and productive skills might be a good first step, just as it might benefit the validity of score interpretations if test components were weighted according to their relative importance in real life. Perhaps these steps would have a positive impact on the number of false negatives. Alternatively, it is

conceivable that students who do not pass the entrance test, yet score above a certain threshold, can still register for university, but with a reduced study load and compulsory language classes. Recently, Ghent University started a similar pilot program, but the outcomes are as yet unknown. Another implication is more general. As it is unlikely that Flanders is the only region where the validity of test score interpretations and the justice of admission procedures can be improved, one could argue for a widespread critical assessment of explicit or implicit claims that underpin test score use in high-stakes contexts.

Limitations

This study was conducted in Flanders, an atypical setting in an Anglo-American-dominated field. Generalizing the results of this study beyond Flanders should be done carefully. Furthermore, the number of participants (55) involved in this study is rather small for quantitative analyses, though substantial for a qualitative study. Still, it is important to keep in mind sample sizes when considering score gains, group differences, and effect sizes. Lastly, the university staff focus groups dealt with the linguistic demands of first-year students, not with the expectations of master's students. Possibly the expectations of L2 students entering Flemish universities at the master's level would have been different. Overall, however, the L2 master's students did not report any concerns that were remarkably different from those expressed by those at the bachelor's level.

Acknowledgements

This study could not have happened without the cooperation of ITNA and STRT. It takes a courageous and self-critical test developer to participate in a study like this. I am also thankful to the reviewers and editors at Language Testing and Language Assessment Quarterly for their helpful and constructive comments, and to Dr. Andries De Smet for his advice on theories of justice. I would also like to express my sincere gratitude to all participants for the time and energy they invested in this project.

A very special word of thanks goes to the $L2_2$ participants. Thank you for allowing me to be a part of your lives during a rough and eventful year. Thank you for your openness, your honesty, and your persistence. Thank you for making me see the value of education and the fullness of the story behind a test score, and for making me realize what high stakes truly mean.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Amuzie, G., & Winke, P. (2009). Changes in language learning beliefs as a result of study abroad. *System*, *37*(3), 366–379.

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

Bachman, L., & Palmer, A. (2010). Language assessment in practice. Oxford: Oxford University Press.

Belzile, J., & Öberg, G. (2012). Where to begin? Grappling with how to use participant interaction in focus group design. *Qualitative Research*, *12*(4), 459–472.

Bérešová, J., Breton, G., Noijons, J., & Szabó, G. (2011). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). Highlights from the Manual*. Strasbourg: Council of Europe.

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement*, *50*(1), 110–114.

Byrnes, H., Maxim, H., & Norris, J. (2010). Realizing advanced foreign language writing. *The Modern Language Journal*, *94*, 1–202.

Carlsen, C. H. (in press). The adequacy of the B2-level as university entrance requirement. *Language Assessment Quarterly*.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421–442.

Clapham, C. (2000). Assessment for academic purposes: Where next? System, *28*(4), 511–521.

Creswell, J. (2015). *A concise introduction to mixed methods research*. Los Angeles, CA: SAGE Publications.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*(1), 1–8.

Davies, A. (2010). Test fairness: A response. *Language Testing*, *27*(2), 171–176.

De Bruyn, K. (2011). *De wet van de sterksten?* Universiteit Gent: Beleidscel Diversiteit en Gender.

De Geest, A., Steemans, S., & Verguts, C. (2015, March). ITNA en ITACE: Twee high-stakestaaltoetsen. Presented at *50 Jaar ILT*, Leuven.

De Standaard. (2013). Selecteer studenten na eerste semester. Retrieved from www.standaard.be

De Wachter, L., & Heeren, J. (2011). Taalvaardig aan de start. Een behoefteanalyse rond taalproblemenen remediëring van eerstejaarsstudenten aan de KU Leuven. Leuven: ILT.

De Wit, K., Van Petegem, P., & De Maeyer, S. (2000). *Gelijke kansen in het Vlaamse onderwijs: het beleid inzake kansengelijkheid*. Leuven & Apeldoorn: Garant.

Dewey, D., Bown, J., Baker, W., Martinsen, R., Gold, C., & Eggett, D. (2014). Language use in six study abroad programs: An exploratory analysis of possible predictors. *Language Learning*, *64*(1), 36–71.

Dey, I. (1993). *Qualitative data analysis*. London: Routledge.

Dworkin, R. (2003). Equality, luck and hierarchy. *Philosophy & Public Affairs*, *31*(2), 190–198.

Dworkin, R. (2011). *Justice for hedgehogs*. Cambridge, MA: Harvard University Press.

Field, J. (2011). Into the mind of the academic listener. *Journal of English for Academic Purposes*, *10*(2), 102–112.

Figueras, N., North, B., Takala, S., Van Avermaet, P., & Verhelst, N. (2009). Relating language examinations *to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A manual*. Strasbourg: Council of Europe.

Fløttum, K., Gedde-Dahl, T., & Kinn, T. (2006). *Academic voices: Across languages and disciplines*. Amsterdam: John Benjamins.

Foucault, M. (1977). *Discipline and punish. The birth of the prison*. London: Penguin.

Freeman, M. (2000). Knocking on doors: On constructing culture. *Qualitative Inquiry*, *6*, 59–369.

Gorin, J. (2007). Reconsidering issues in validity theory. *Educational Researcher*, *36*(8), 456–462.

Gilabert, R. (2005). Evaluating the use of multiple sources and methods in needs analysis: A case study of journalists in the Autonomous Community of Catalonia (Spain). In M. Long (Ed.), *Second language needs analysis* (pp. 182–200). Cambridge: Cambridge University Press.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine de Gruyter.

Hulstijn, J. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*(3), 229–249.

Hulstijn, J. (2014). The Common European Framework of Reference for Languages. A challenge for applied linguistics. *International Journal of Applied Linguistics*, *165*(1), 3–18.

Hume, D. (1978). *A treatise of human nature*. Oxford: Clarendon Press.

Hyland, K., & Hamp-Lyons, L. (2002). EAP: Issues and directions. *Journal of English for Academic Purposes*, *1*(1), 1–12.

ILTA. (2000). ILTA code of ethics. Retrieved from www.iltaonline.com

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kane, M., Kane, J., & Clauser, B. (2017). A validation framework for credentialing tests. In C. Buckendahl & S. Davis-Becker (Eds.), *Testing in the professions: Credentialing Polices and Practice* (pp. 20–41). New York: Routledge.

Kinginger, C. (2008). Language learning in study abroad: Case studies of Americans in France. *The Modern Language Journal*, *92*, 1–124.

Leuven, KU. (2015). *Onderwijs- en examenreglement 2015-2016*. Retrieved from www.kuleuven.be

Kunnan, A. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge: Cambridge University Press.

Kunnan, A. (2007). Introduction: Test fairness, test bias and DIF. *Language Assessment Quarterly*, *4*(2), 109–112.

Kunnan, A. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, *27*(2), 183–189.

Lado, R. (1961). *Language testing*. London: Longman.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lee, Y., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research*, *1*(4), 366–389.

Lievens, S. (2016). *Diversiteit aan de UGent: de instroom van kansengroepen in cijfers*. Ghent: Universiteit Gent.

Lissitz, R., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448.

Llanes, À., Tragant, E., & Serrano, R. (2012). The role of individual differences in a study abroad experience: The case of Erasmus students. *International Journal of Multilingualism*, *9*(3), 318–342.

Long, M. (2005). Methodological issues in learner needs analysis. In M. Long (Ed.), *Second language needs analysis* (pp. 19–79). Cambridge: Cambridge University Press.

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes*, *10*(2), 79–88.

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.

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

Miles, M., & Huberman, A. (1994). *Qualitative data analysis*. Beverly Hills, CA: SAGE Publications.

Morita, N. (2004). Negotiating participation and identity in second language academic communities. *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect*, *38*(4), 573–603.

Nederlandse Taalunie (2015). *Totstandkoming*. Retrieved from http://taalunieversum.org, 20 April 2017.

Oller, J. (2012). Grounding the argument-based framework for validating score interpretations and uses. *Language Testing*, *29*(1), 29–36.

O'Sullivan, B. (2016). A story to tell, a lesson to learn: The testing industry and validation. Presented at the ALTE 48th Conference, Stockholm.

Patton, M. (2002). *Qualitative evaluation and research methods*. Newbury Park, CA: SAGE Publications.

Phillips, D. (2007). Adding complexity: Philosophical perspectives on the relationship between evidence and policy. *Yearbook of the National Society for the Study of Education*, *106*(1), 376–402.

Rawls, J. (1971). A theory of justice. Cambridge, MA: Harvard University Press.

Rawls, J. (2001). *Justice as fairness: A restatement*. Cambridge: Belknap Press.

Reybold, L., Lammert, J., & Stribling, S. (2013). Participant selection as a conscious research method: Thinking forward and the deliberation of "emergent" findings. *Qualitative Research*, *13*(6), 699–716.

Römhild, A., Kenyon, D., & MacGregor, D. (2011). Exploring domain-general and domain-specific linguistic knowledge in the assessment of academic english language proficiency. *Language Assessment Quarterly*, *8*(3), 213–228.

Sen, A. (2010). *The idea of justice*. London: Penguin.

Serrano, R., Tragant, E., & Llanes, À. (2012). A longitudinal analysis of the effects of one year abroad. *Canadian Modern Language Review*, *68*(2), 138–163.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Harlow, UK: Longman.

Snow, C. (2010). Academic language and the challenge of reading for learning about science. *Science*, *328*(5977), 450–452.

Smet, P. (2011). *Samen taalgrenzen verleggen*. Brussels: Departement Onderwijs.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes*, *10*(2), 89–101.

Turner, C. (2014). Mixed methods research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–15). New York: John Wiley & Sons.

Toulmin, S. E. (2003). *The uses of argument*. Rev. edn. Cambridge, UK: Cambridge University Press.

Universiteit Antwerpen (2015). *Procedure proc/adond/001.1*. Retrieved from www.uantwerpen.be

Universiteit Gent. (2015). *Onderwijs- en examenreglement 2015-2016*. Retrieved from www.ugent.be

Universiteit Hasselt. (2015) *Toelatingsvoorwaarden*. Retrieved from www.uhasselt.be

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing*, *11*(3), 321–344.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170.

Xi, X., Bridgeman, B., & Wendler, C. (2013). Tests of English for academic purposes in university admissions. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 318–337). New York: John Wiley & Sons.

Appendix 1. ITNA and STRT: Tasks and criteria.

	ITNA		STRT	
Writing	Reading: 5 texts, multiple-choice questions	Binary (selected response)	Summarizing a scientific article	Ct V G C S
	Reading: restructure jumbled paragraphs		Taking notes based on a scripted lecture	
	Listening: 3 recordings, multiple-choice questions		Argumentative writing based on written input	Ct V G C S R
	Listening: fill-in-the-gaps		Argumentative writing based audio input	
	Language-in-use: cloze (vocabulary)			
	Language-in-use: cloze (vocabulary)			
	Language-in-use: grammar			
	Language-in-use: fill-in-the-gaps			
	Language-in-use: multiple choice (Vocabulary)			
	Giving a presentation, based on graphic input			
	Argumentative speaking, based on graph or table			
Speaking	Giving a presentation, based on graphic input	V G C F P	Giving a presentation, based on graphic input	Ct V G C P F I R
	Argumentative speaking, based on graph or table		Argumentative speaking, based on graph or table	

Note: C = Cohesion, Ct = Content, F = Fluency, G = Grammar, I = initiative, P = Pronunciation, R = Register, S = Spelling, V = Vocabulary.

Appendix 2. $L2_1$ participants.

Faculty	B/M[a]	M/F[b]	L1	Nationality	L2[c]	Test[d]	ID
Engineering	B	F	German	Switzerland	24	STRT	S1
Medicine	M	F	Farsi	Iran	17	ITNA	S2
Law	M	M	Turkish	Turkey	9	ITNA	S3
Sciences (Geology)	M	M	French	France	12	ITNA	S4
Political sciences	M	F	Greek	Greece	48	ITNA	S5
Law	M	F	French	Belgium	12	ITNA	S6
Social sciences	B	M	English	South Africa	36	ITNA	S7
Psychology	M	F	Finnish	Finland	14	ITNA	S8
Sciences (Chemistry)	M	M	Arabic	Morocco	12	ITNA	S9
Political sciences	M	F	Turkish	Turkey	36	ITNA	S10
Medicine	B	F	English	Cameroon	12	ITNA	S11

[a]Bachelor/master.
[b]Male/female.
[c]Months of Dutch L2 instruction.
[d]University entrance test taken.

Appendix 3. $L2_2$ participants.

Faculty	U[a]	B/M[b]	M/F[c]	L1	Nationality	L2[d]	STRT	ITNA	+/−[e]	ID
Engineering	G	M	F	Ukrainian	Ukraine	18	1	1	+	S12
Economics	G	M	F	Spanish	Peru	7	1	1	+	S13
Law	G	M	F	French	Belgium	120	1	1	+	S14
Political sciences	I	M	F	Haitian	Haiti	20	1	0	+	S15
Psychology	G	B	F	Spanish	El Salvador	12	1	0	+	S16
History	L	M	F	Turkish	Turkey	12	1	1	+	S17
Linguistics	L	B	F	Ukrainian	Ukraine	9	1	1	+	S18
Law	A	B	F	Albanian	Albania	8	1	1	+	S19
Engineering	G	B	M	Farsi	Iran	6	0	1	−	S20
Biomedical	L	M	F	French	Congo	10	1	0	−	S21
Linguistics	L	B	F	Spanish	Costa Rica	11	1	0	−	S22
Psychology	L	B	F	German	Germany	10	1	1	−	S23
Psychology	L	B	M	Vietnamese	Vietnam	12	1	1	−	S24
Engineering	L	M	F	Russian	Russia	22	1	1	−	S25
Law	A	M	F	French	Belgium	72	1	1	−	S26
Law	A	M	F	French	Belgium	72	1	1	−	S27
Economics	H	B	F	Armenian	Armenia	7	0	0	V	S28
Economics	A	B	M	Pashto	Afghanistan	14	1	0	V	S29
Medicine	G	M	F	Spanish	Chile	24	1	0	?	S30
Economics	L	B	F	French	Belgium	72	1	1	?	S31

[a]Ghent University/Inter-university degree/University of Leuven/Antwerp University/University of Hasselt.
[b]Bachelor/master.
[c]Male/female.
[d]Months of Dutch L2 instruction.
[e]more (+) or less (-) than 50% of courses passed / Attrition due to Visa or immigration issues / Reason for attrition unknown (?).

Appendix 4. University staff.

Faculty/department	Position	ID
Central administration	Didactics policy manager	Ac2
	University director of educational affairs	Ac12
	Language policy manager	Ac5
	Language policy manager	Ac10
Humanities	Professor	Ac1
	Tutor	Ac16
	Professor	Ac17
	Faculty director of educational affairs	Ac22
	Tutor	Ac15
	Professor	Ac7
Engineering	Tutor	Ac11
	Professor	Ac6
	Faculty director of educational affairs	Ac23
Medicine	Faculty director of educational affairs	Ac14
Sciences	Professor	Ac13
	Faculty director of educational affairs	Ac18
	Professor	Ac20
Economics	Tutor	Ac3
	Faculty director of educational affairs	Ac21
Law	Professor	Ac8
	Tutor	Ac19
Psychology	Faculty director of educational affairs	Ac9
Social and political sciences	Professor	Ac4
	Tutor	Ac24