Bart Deygers Assessing high-stakes assumptions

ASSESSING HIGH-STAKES ASSUMPTIONS

Bart Deygers

Supervisor Prof. dr. Kris Van den Branden
Co-supervisor dr. Koen Van Gorp

ISBN: 9789 079219070

ASSESSING HIGH-STAKES ASSUMPTIONS

A longitudinal mixed-methods study of university entrance language tests, and of the policy that relies on them.

Proefschrift ingediend tot het behalen van de graad van Doctor in de Taalkunde door Bart Deygers, 2017

Promotor
Prof. dr. Kris Van den Branden
Copromotor
dr. Koen Van Gorp

Begeleidingscommissie
Prof. dr. Lieve De Wachter
Prof. dr. John Norris

Juryleden
Prof. dr. Claudia Harsch
Prof. dr. Elke Peters

KU LEUVEN

Centrum voor Taal nt2 en Onderwijs

KU Leuven
Faculteit Letteren
Onderzoekseenheid Taalkunde
Onderzoeksgroep Taal & Onderwijs

ISBN: 9789079219070
Printed by ACCO Drukkerij – Herent, Belgium

This work was funded by the Research Foundation Flanders, FWO Vlaanderen, under Grant number G078113N.

For Max

I wake with the sparrows
And I hurry off to work
The need for validation, baby
Gone completely berserk

Nick Cave

A PERSONAL NOTE

This PhD started out as a project proposal – an unreadable title and an imaginary timeline. Gradually, it stopped being a project, and started to truly matter. It's taught me a lot, sharpened my mind and deepened my thinking. I will miss it, and I will always be grateful to the people who helped shape it. This is for you.

Thank you, Kris for the open discussions, and the right kind of encouragement. For giving me freedom and trust, for asking thought-provoking questions, and for answering my outbursts with a pinch of stoicism. Thank you, Koen for giving me so many opportunities. For encouraging me to do research in the first place. For countless beers on three continents, and for a bunch of memorable late nights. Thank you both for the extra efforts you made during the final weeks of writing. John. You changed everything when you invited me to Georgetown. Thank you for looking out for me, for your to-the-point comments, and for making my 35th birthday so special. Thank you, Lieve, for your moral support, your constructive comments, and your spontaneity. Thank you Elke, for the coffees and the solid advice. And Claudia, for combining inspiring research with societal involvement.

Cecilie, sparring partner extraordinaire, thank you so much for the animated discussions, the solid advice, the personal talks, the unstoppable laughs and the unwavering support. To all my colleagues at the Centre for Language and Education: thank you! A big shout-out to Kathelijne, Joke, Goedele, Lies, Carolien for the uplifting talks, and to the CNaVT crew for the help. This study could not have happened without your cooperation, and without the assistance of ITNA. It takes a courageous and self-critical test developer to participate in a study like this. Thank you, Young A, Sandra, John, Tyler, Meg, Mina, Todd, Amy, and Francesca – my dear Georgetown friends – for setting some impressively high standards. Thank you, Lourdes, for helping me get Chapter 5 to the next level.

To get this research off the ground, quite a few people helped out. Bart, Caroline, Carolien, Delphine, Sofie, Eefje, Ellen, Freek, Inge, Jordi, Kathelijne, Martien, Nele, Sara, Sarah, Sibo, Sien, Lisa, Jackie, Eva, Vanessa, Willemijn, mum, dad, and Marie-Paule – you guys are magnificent. And thank you Nick, Mariangela, Esther, Beate, Dina, and the other wonderful people in ALTE. Chapter 1 would have looked very differently without your help,

Thank you Ellen. For the talks, the openness, the laughs and the warmth. For the weird food we eat. And Tom, for leading by example when it comes to doing a PhD. 130 represent. Thank you, Sibo for challenging my assumptions about the world, Sünbül for your unique blend of frankness, and Marieke for fooling me annually. Thank you, Jordi and Filip for the best kind of peer pressure. Thomas

and Laure-Ann, for the scientific and moral support, Sara and Aline for the think tank. Kris, for the no-frills encouragement, and Sara for a stellar C1 performance. Andries, for discussing John Rawls on a random train ride. Sarah, for our wide-ranging lunchtime discussions, and for your take on justice and policy. Also, Mandy. Nobody has been at my side quite as much as you have during these past years. Thank you Sheila, Bernadette and Suzanne, for the oinks and the pokes.

Thank you for the love of language and education, and the help along the way mama, papa, Veerle and Katrien. Sofie! I can't thank you enough. For reminding me of the bigger picture, for pushing me to explore life to the fullest. For telling me not to cut corners, and for defusing the stressful moments that invariably occur in a project like this with disarming jokes, wisdom, and proofreading wizardry.

A very special word of thanks goes to the respondents. I am indebted to the academic respondents and the policy makers, for their candor and participation. I owe the L2 students who took part in the longitudinal study a world of gratitude. Thank you for allowing me to be a part of your lives during a rough and eventful year. Thank you for making me see the full value of education and the story behind a score. Thank you for making me realize what high stakes truly mean.

CONTENTS

Introduction: Examining assumptions **12**
- The university entrance policy in Flanders, Belgium 12
- Validating test score use 18
- A note on policy 21
- Research goals 22
- Available evidence 27
- Research design 29
- Structure and relevance of this dissertation 34
- A note on composition 36

PART 1: LEVELS & CONSTRUCTS **38**

Chapter 1: University admission policies across Europe **40**
- Research questions 42
- Participants & methodology 42
- Results 45
- Discussion 51
- Conclusion 52

Chapter 2: Content & level representativeness **54**
- Academic language requirements 54
- Justice 55
- Research aims 57
- Participants & methodology 58
- Results 64
- Discussion 74
- Conclusion 76

PART 2: SELECTION & DISCRIMINATION **78**

Chapter 3: Level & construct equivalence **80**
- Equivalence 80
- Justice 82
- Research questions 83
- Participants & methodology 83
- Results 88
- Discussion 94
- Conclusion 97

Chapter 4: Criterion equivalence **100**
Criterion equivalence and the CEFR 101
Research question 103
Participants & methodology 103
Results 106
Discussion 112
Conclusion 113

Chapter 5: Comparing L1 and L2 performance **116**
Research into L1 and L2 performance 116
Research questions 120
Participants & methodology 121
Results 127
Discussion 131
Conclusion 134

PART 3: GAINS & CONTEXT **136**

Chapter 6: Examining L2 gains **138**
Research questions 143
Participants & methodology 145
Results 149
Discussion 168
Epilogue 173

PART 4: POLICY, CONCLUSION & IMPLICATIONS **176**

Chapter 7: The policy-making process **178**
Examining university admission policies 178
Research questions 180
Participants & methodology 180
Results 182
Discussion 186
Conclusion 187

Chapter 8: Summary & discussion of the research findings **190**
Research goal 1 190
Research goal 2 194
Research goal 3 199
Research goal 4 200
Research goal 5 202

Chapter 9: Limitations, implications & recommendations **204**
A few words on strengths and limitations 204
Implications 206
Recommendations 211
"The need for validation, baby, gone completely berserk" 214

References 216

Academic output related to this PhD 239

Summary in Dutch 242

Appendix 247

TABLES & FIGURES

Introduction

Table 1.1. International students 13
Table 1.2. Language requirements 15
Table 1.3. Double coding 33
Figure 1.1. Toulmin's argument structure 19
Figure 1.2 Validation scheme 20
Figure 1.3 Research design 30

Chapter 1: University admission policies across Europe

Table 2.1. Countries and regions surveyed 43
Table 2.2. Tests accepted for university entry 45
Table 2.3. CEFR level required for university entrance 46
Table 2.4. Is B2 enough 46
Table 2.5. Who decides on language requirements 47
Table 2.6. Empirical research 48

Chapter 2: STRT & ITNA: Content & level representativeness

Table 3.1. Focus group samples, arranged by CEFR level 62
Table 3.2. University lectures and STRT listening prompts 66
Table 3.3. Academic language skills selected in focus groups 71

Chapter 3: level & construct equivalence

Table 4.1. Research vs. regular population 85
Table 4.2. Descriptive statistics 89
Table 4.3. Pass/Fail crosstab 89
Table 4.4. ITNA computer scores & STRT written scores 90
Table 4.5. MFRA written tasks (equal weights) 91
Table 4.6. Promax-rotated factor loadings 92
Table 4.7. MFRA oral criteria (equal weights) 93
Table 4.8. MFRA oral criteria (actual weights) 94

Chapter 4: criterion equivalence

Table 5.1. Jaccard index for rating descriptor pairs 107
Table 5.2. Frequencies of assigned CEFR levels 108
Table 5.3. Probability of attaining B2 or higher 108
Table 5.4. Relationship between corresponding criteria 109
Table 5.5. Multivariate linear regression 110
Table 5.6. Linear regression on criterion level 111
Table 5.7. MFRA: STRT and ITNA (by measure) 111
Table 5.8. MFRA: STRT and ITNA criteria (by measure) 112

Chapter 5: Comparing L1 and L2 performance

Table 6.1. Multivariate linear regression 122
Table 6.2. Demographic data of participants 123
Table 6.3. Promax rotated factor loadings 126
Table 6.4. Descriptive statistics 128
Table 6.5. MFRA for facet "Group" 128
Table 6.6. Wilcoxon signed-rank test: Flemish, $L2_F$ and $L2_I$ 129
Table 6.7. Wilcoxon signed-rank test: L1, G1.5, and $L2_F$ 129
Table 6.8. Logistic regression: Flemish ~ criterion scores 130
Table 6.9 Multinomial linear regression 131

Chapter 6: examining L2 gains

Table 7.1. Multivariate linear regression 147
Table 7.2 Language gains 150
Table 7.3. Reduced data matrix: interpersonal 151
Table 7.4 interpersonal relationships 152
Table 7.5. How do think your classmates see you? (March) 158
Table 7.6. Reduced data matrix: institution 161

Chapter 7: the policy-making process

Table 8.1. Policy maker respondent codes 180
Table 8.2. Data coding categories 181
Table 8.3. Exemptions from admission requirements 183

Chapter 8: summary & discussion of the research findings

Table 9.1. STRT & ITNA task types 195
Table 9.2. STRT & ITNA result vs. academic success 198

APPENDICES

Appendix 1 (1/3). STRT Part 1: Listening-into-writing 247
Appendix 1 (2/3). STRT, Part 2: Reading-into-writing 248
Appendix 1 (3/3). STRT, Part 3: Speaking 249

Appendix 2 (1/2). ITNA: computer test 250
Appendix 2 (2/2). ITNA: Speaking test 251

Appendix 3. L2P participants 252

Appendix 4. L2F participants 253

Appendix 5. University staff 254

INTRODUCTION
EXAMINING ASSUMPTIONS

If everybody has the right to an education (UN General Assembly, 1948), and if everybody has the right to pursue the goals that he or she deems valuable (Nussbaum, 2002; Sen, 2010), then an entrance policy that obstructs people's access to the education of their choosing by means of an assessment procedure would require strong empirical justification (Sen, 2010). It would have to be clear that the admission policy facilitates the selection of the right applicants for the right reasons. Nevertheless, in many contexts, university entrance requirements are based on assumptions or claims that are as yet unsupported by empirical data (McNamara & Ryan, 2011).

The primary assumption supporting the use of language tests to control entrance to higher education is that a certain language proficiency level is required to verify whether L2 students will be able to meet the linguistic requirements of academic studies. This assumption is rarely investigated or challenged, however, even though its impact is substantial (McNamara & Ryan, 2011). The purpose of this dissertation is to empirically assess the assumptions that support the use of language tests in one specific case: the Flemish university entrance policy. When test scores are used in such a way that they may fundamentally impact a test-taker's opportunities in life, the stakes are high. Such tests, and the claims that are made on the basis of their scores, require close scrutiny and robust evidence.

THE UNIVERSITY ENTRANCE POLICY IN FLANDERS, BELGIUM

For most students who have graduated from a Flemish high school, there are no obligatory or binding university entrance tests. Only Flemish students who wish to pursue a degree in medicine or dentistry need to pass an examination that tests their knowledge in exact sciences, and their reading skills. For all other students with a Dutch high school degree there are no centralized subject-specific tests or language tests prior to university entrance.

The relatively open university entrance policy has resulted in large groups of students in the first year, where ex cathedra teaching (i.e., one-way transmission teaching) is the norm. Consequently, students are generally not expected to speak or write much in the course of their studies until the second or the third year, when the first written papers are due (De Wachter, Heeren, Marx, & Huyghe, 2013). Another consequence of the open registration policy is that the de facto selection of students occurs not before the start of academic programs at

university but at the end of the first year, when around 60% of the students fail their exams (Amkreutz, 2013). Students with an atypical educational background, a low socio-economic status, or an L1 different from Dutch are overrepresented in the group of students who do not pass their first year at university (De Wit, Van Petegem, & De Maeyer, 2000; Lievens, 2016).

The number of international students at Flemish universities has been steadily increasing in recent years (Beleidscel Diversiteit en Gender, 2016), even though the proportion of international students at Flemish universities is still considerably lower than at their British and North American counterparts.

Table 1.1. International students at Flemish universities (2015-'16)

	Student population	International students
University of Leuven	41.500	19%
Ghent University	41.000	11%
University of Antwerp	20.000	14%
University of Brussels	11.000	20%
University of Hasselt	6.000	9%

Note. These numbers include PhD students

International students account for up to 20% of the population at the five Flemish universities (see Table 1.1). These publicly available percentages also include PhD students, however, who are not required to attend curricular classes and who do not need to pass Dutch language tests. The proportion of international L2 students who are required to pass a language test (i.e., those at the bachelor or master level) is considerably lower than the publicly available numbers, but few universities are willing to disseminate detailed information about the actual composition of the non-PhD student population. At Ghent University, 1.7% of the newly registered freshmen in 2015 were international students (private communication). Importantly, universities often count Dutch-speaking students from the Netherlands – who are exempt from taking a language test – as international students. The University of Antwerp does keep track of the proportion of international L2 students at the undergraduate level. At this institution, less than 4% of the undergraduate student population consists of international L2 students (private communication), which is substantially below the publicly available figure of 14%. Given the disparity in definitions and inclusion criteria it is impossible to state how many international L2 students are impacted by the university entrance policy.

Language regulations

It has been argued that the educational language policy in Flanders is to a large extent based on ideology (Blommaert, 2011; Blommaert & Van Avermaet, 2008; Van Splunder, 2015). Policy makers do not always favor increased internationalization of the student population. In the minds of many, higher education is still primarily seen as a service to Flemish taxpayers (Leliaert, 2011; Truyts & Torfs, 2015). Congruently, the language of education in most programs is Dutch, the official language of Flanders. As a result of nearly two centuries of language-related political turmoil an ideology of territorial monolingualism (Blommaert, 2011; Van Splunder, 2015) has permeated many aspects of Flemish society, including education. Many primary and secondary schools use a strict Dutch-only policy (Agirdag, 2010; Blommaert & Van Avermaet, 2008; Strobbe, 2016) and the official language policy of higher education in Flanders has been influenced by the same ideology.

The official language of Flemish higher education is Dutch, in administrative and educational matters (Vlaamse Regering, 2013). The governmental decrees that shape language regulations in Flemish higher education strive towards maintaining Dutch as an academically viable language. Recent rulings have become more lenient towards organizing education in languages other than Dutch, but can still be considered rather restrictive when compared to The Netherlands *(Wet op het hoger onderwijs en wetenschappelijk onderzoek, 1992)*, where over 60% of the bachelor and maser programs were taught in English in 2016 (Bouma, 2016). A Flemish university cannot organize more than 6% of its bachelor or 35% of its master programs in a language other than Dutch. A program is considered non-Dutch when more than 18.33% (bachelor programs) or 50% (master programs) of the classes are not taught in Dutch (Departement Onderwijs en Vorming, 2015).

In contrast to their peers who graduated from a Dutch-medium high school, international L2 students in Flanders need to prove a certain language proficiency level. The level of language proficiency required by all Flemish universities for bachelor and master programs, is the B2 level of the *Common European Framework of Reference for Languages* (CEFR - Council of Europe, 2001).

The B2 level is the fourth of six consecutive language proficiency levels on the CEFR (Council of Europe 2001), which starts at A1 and goes up to the very advanced C2 level. In the following chapters the B2 level, as well as its applications and applicability in university entrance testing, will be discussed in detail. For now, it suffices to describe the B2 learner as somebody with a language proficiency that is comparable to ACTFL Advanced Mid (ACTFL, 2016), 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 argumentation (Council of Europe, 2001: 24).

The CEFR has been widely adopted by educational policy makers, and its levels are used to determine entrance requirements in a wide variety of contexts (Figueras 2012). Unfortunately, however, the CEFR bands are rather broad (Fulcher 2004; Hulstijn 2014), and two tests that link to the same level are not necessarily equally difficult, even though policy may assume that they are (Green, Forthcoming). Because there is substantial room for variation within one and the same CEFR level, it is important not to assume equivalence of tests simply because they share the same CEFR level, without empirically investigating that assumption (see the special issue of *The Modern Language Journal* edited by Byrnes, 2007).

International L2 students who wish to pursue a program in which Dutch is not only the medium of instruction, but also the goal (i.e., translation studies, or Dutch literature) are required to prove C1 proficiency at some universities (e.g., Ghent University, and University of Antwerp as of 2017), while others require C1 for teacher training programs (e.g., University of Hasselt). B2, however, can be considered the default entrance requirement (see Chapter 6).

Table 1.2. Language requirements for international L2 undergraduate students at Flemish universities

	Level	Accepted proof of language proficiency								
	B2 requirement	STRT & ITNA[6]	60 credits in Flemish HE[7]	Flemish SE[8] degree	STEX II[9]	One year in Flemish SE	Accredited B2 course	Medicine/dentistry admission test	HOSP[10] degree	Limburg HE language test
University of Leuven[1]	★	★	★	★			★			
Ghent University[2]	★	★	★	★	★	★	★	★		
University of Antwerp[3]	★	★	★	★	★	★				
University of Brussels[4]	★	★	★	★	★	★		★	★	
University of Hasselt[5]	★	★	★	★	★	★				★

Note. [1] KU Leuven, 2016, p. 4; [2] Universiteit Gent, 2016, p. 19; [3] Universiteit Antwerpen, 2016, p. 3; [4] Vrije Universiteit Brussel, 2014, p. 22, [5] Universiteit Hasselt, 2016

[6] ITNA: *Interuniversitaire Taaltoets Nederlands voor Anderstaligen* (Inter University test of L2 Dutch), STRT: *Educatief Startbekwaam* (Ready to start higher education), [7]Higher Education, [8]Secondary Education, [9]*Staatsexamen Programma II* (State Exam, Netherlands), [10]*Hoger Onderwijs voor Sociale Promotie* (higher education for social promotion)

Different universities allow for different kinds of evidence but certain documents are accepted at all five universities as adequate evidence of B2 ability (Table 1.2). A certificate of STRT or ITNA, two accredited B2 tests (see below), grants admission to every Dutch-medium program. Similarly, a degree of a Dutch-medium high school or sixty credits in a Flemish higher education program are considered as sufficient proof of B2 ability.

The B2 requirement is not imposed on international students only; international teaching staff too need to prove B2 ability if they do not teach in Dutch, C1 if they do (Vlaamse Regering, 2013). Within three years after being appointed they are required to pass ITNA (Departement Onderwijs en Vorming, 2015).

STRT & ITNA

ITNA (*Inter University test of L2 Dutch*) is a computer-based and face-to-face test, issued and developed by the *Interuniversitair Testing Consortium* of Flemish university language centers. STRT (*Ready to start higher education*), a task-based, integrated-skill language test, is the only Dutch language test at the B2 level that is internationally administered. It is developed at the University of Leuven, and funded by the Dutch Language Union, an international, intergovernmental organization overseeing the Dutch language policy in the Netherlands, Belgium and Suriname.

The certificates of these two tests are accepted by all Flemish universities as proof of the required B2 ability. Both tests have been linked to the B2 level of the CEFR (Council of Europe, 2001) following the *familiarization*, *specification*, *standard setting and validation* procedures described in Figueras, North, Takala, Verhelst, & Van Avermaet (2005).

STRT and ITNA are comparable on a number of parameters other than their CEFR level. Both tests have undergone a successful audit by the *Association of Language Testers in Europe* (ALTE), offering an independent assessment of their validity, reliability, and consistency. They also share the same primary purpose (i.e., testing non-native speakers of Dutch for university admission), and refer to communicatively oriented conceptualizations of language proficiency, such as Weir's (2005) sociocognitive framework and the CEFR, as primary sources of their construct. Finally, since both tests employ a pass/fail procedure, candidates either attain a B2 certificate, or not. In terms of operationalization STRT and ITNA show a number of substantial differences, especially in the written component. The oral components are more comparable, although there are differences in the rating criteria used. Appendix 1 and 2 give an overview of the operationalization of STRT and ITNA respectively.

The written component of STRT is paper-based and consists of two components. In the writing-from-listening component candidates write an argumentative text based on audio input, and summarize a scripted lecture about a general topic. The writing-from-reading component also includes an argumentative task, and a summary task with a substantial argumentative component. The written component of ITNA is computer-based and primarily includes selected-response question types. Candidates drag jumbled paragraphs to order them correctly, fill out missing words in a text, and answer multiple-choice questions about reading or listening prompts. At the B2 level, ITNA does not include writing tasks, as writing is measured indirectly using selected-response item types.

The oral components of both tests do not take more than 25 minutes, including preparation time. Candidates interact with a trained examiner during the oral component, which consists of a presentation and an argumentation task. The argumentation task invites the test takers to weigh a number of alternative solutions to a problem, and argue why their choice is the better one. In the presentation task candidates briefly present a study by using input material such as graphs and tables. Even though the oral tasks are similar in both tests, the scoring rubrics differ, because ITNA only takes into account linguistic criteria (*Vocabulary*, *Grammar, Cohesion, Pronunciation, Fluency*), while STRT also focuses on content (i.e., whether a performance contains the main points mentioned in the prompt).

STRT writing tasks are rated by two independent trained raters who score content criteria in a binary way (i.e., whether the candidate mentions the required aspects or not) and linguistic criteria on a four-point scale. The ITNA computer test is scored electronically using a binary rating procedure, while its oral component is scored in situ by the examiner and an additional rater, who come to a joint overall score for five linguistic criteria. The oral STRT component is administered by a trained examiner, recorded and centrally scored by two independent trained raters, using a rating scale that includes five criteria that correspond to those used in ITNA, plus *Register*, *Initiative* and *Content* (i.e., whether the candidate mentions all salient points asked for in the prompt). ITNA examiners and raters tend to be experienced L2 teachers of Dutch who typically attend training at least once a year and score oral tests at different times throughout the year. STRT raters are usually novice raters with a background in linguistics or communication studies who have received a two-day training and have shown that they are able to rate two sample exams with a satisfactory degree of accuracy and consistency. The first day of training includes becoming acquainted with the test, its purpose and the rating scales. After this, candidate raters score eight standardized performances of each task (i.e., 48 performances), comparing their score to the standardized one. At the end of the second day of the training, candidate raters score two exams (i.e., 12 tasks). They are considered fit for rating when the scores of all raters correspond to the standardized scores.

STRT candidates do not typically receive a detailed score report that lists scores on task or criterion level. Instead, it provides a certificate for candidates who achieved an overall Rasch measure at or above 1.42 (cut score based on standard setting; private communication, 6 January 2016). ITNA informs candidates whether or not they passed the computer test via e-mail on the same day as the administration. Candidates get their results on each section of the computer test, and those who reached the cut score of 54% are invited to take the oral exam. After the oral component the overall cut score required to obtain ITNA certification is 52.5% (private communication, 6 March 2015).

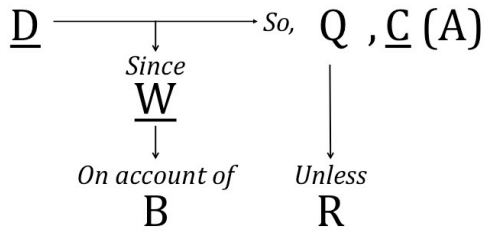
In line with the rising number of international students at Flemish universities, the candidature of STRT and ITNA has been growing. The number of ITNA test takers has risen from 286 in 2010 (the year of the first ITNA administration) to over 1000 in 2014 (Interuniversitair Testing Consortium, 2015). STRT served 608 candidates in 2010, and 957 in 2014 (CNaVT, 2013, 2016b).

VALIDATING TEST SCORE USE

The approach to validation adopted in this dissertation relies on Kane's Toulmin-inspired treatment of explicit and implicit claims concerning the interpretation and use of test scores (Kane, 2013). Throughout the chapters, specific implications and applications of Kane's approach will be explained, but at this point it is useful to discuss the more general framework that links the different chapters.

Test scores bear 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, 2012, 2013). 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. The responsibility of validation therefore does not fall on the test developer alone, but also on the score user. Flawed tests preclude valid score use, and as such the test developer is responsible for developing an instrument of measurement that is valid for the purpose for which it was intended (Norris, 2008). Score users bear the responsibility for proving any additional claims they may make (Kane, 2013).

Figure 1.1. Toulmin's argument structure



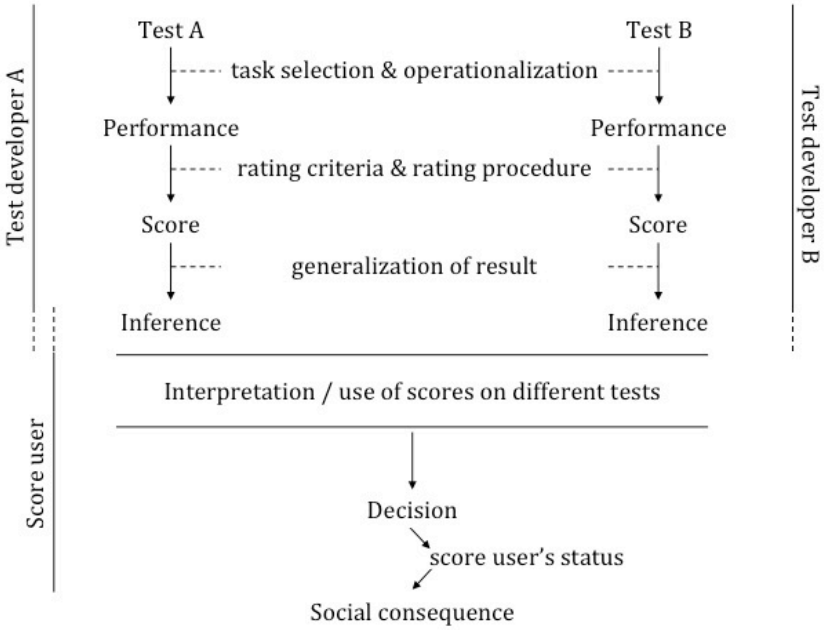
To substantiate claims related to score use, Kane proposes using Toulmin's argument structure (Figure 1.1), which provides a transparent connection between claims and the data on which these claims rely. In Toulmin's logic, a *Claim* (C) is a conclusion based on *Data* (D). Any claim that is unsupported by data, is empty (Toulmin, 2003). Throughout this dissertation, we will examine claims that support the university entrance policy in Flanders. In policy, it is not uncommon for claims to remain implicit (Phillips, 2007), and for that reason the term Assumption (A) will be used throughout this dissertation to refer to policy claims. An assumption, as defined by the Oxford Dictionary refers to "a thing that is accepted as true or as certain to happen, without proof". Typically, data need to be interpreted before they can lead to a claim. This connection between data and the claim is called a *Warrant* (W). *Backing* (B) may be required to offer credible support to the interpretation of the data (e.g., a publication that supports a certain use of the data). Finally, *Qualifiers* (Q) and *Rebuttals* (R) are used to specify the degree to which something may be true (Q), and the conditions under which the claim may not apply (R).

Figure 1.2 offers a visual representation of the validation model employed in this study. The model shows the different stages involved in using more than one language test in a high-stakes language testing policy. The upper half concerns claims that test developers need to prove, whereas the lower half deals with uses and interpretations that are the responsibility of score users.

The basic structure of the model's upper half reminds of Bachman & Palmer's (2010) Assessment Use Argument, and shows how the different steps in the testing process affect the outcome. A candidate's performance on a test is mediated by the selection and the operationalization of test tasks (Bachman, 2002; Sasayama, 2016; Weir, 2005). This performance is then translated into a score by means of a rating process, which may include score transformation by means of statistical procedures, but always involves an interpretative component, realized by human raters or preprogrammed in rating algorithms (Lumley, 2002). Often, but not always, the test developer will then make a score-based assumption concerning a candidate's real-life ability, or link the score to an external framework such as the CEFR. Test developers do not necessarily make

claims about a candidate's real-life language ability, however, and may leave the matter of making inferences to the score user (hence the dotted vertical lines in the model).

Figure 1.2. Validation scheme



Since a test score is influenced by a number of interconnected variables, test developers need to justify a certain amount of implicit or explicit claims relating to the test's level, goal, or population (Kane, 2006; Norris, 2008). At a minimum it needs to be clear that the selected tasks are representative for the target domain (Weigle & Malone, 2016; Weir, 2005), that sufficient measures have been taken to reduce the influence of bias and construct irrelevant variance on test scores (Shaw & Imam, 2013), that scores are assigned in a reliable way, and that there is backing for the inference which links a score to real-life ability or to an external criterion (see the various codes of practice: ALTE, 2001; EALTA, 2000; ILTA, 2007). Throughout the process of validation, the purpose for which the test was designed – the intended use – needs to be clear (Kane, 2001; Norris, 2008). If this precondition is not met, the representativeness of the tasks, the adequacy of the criteria, and the quality of the inference would be very difficult to assess. Determining how fit-for-purpose a test is, seems rater futile if the purpose of the test in unclear. If a test user would decided to use a score for a purpose not intended by the developer, the score user would need to provide evidence that using the test for an originally unintended purpose is warranted (Kane, 2013). Often, policy makers or admission officers at the institutional or at the national level will determine the social consequences of a test score. If this is the case, score users should assume part of the responsibility for validating claims that

impact real-world consequences (Kane, 2013). If an admission board accepts different scores on different tests for the same purpose (e.g., university admission) score users will be required to provide evidence to support this policy.

The final stage of the validation framework used in this study concerns the decision made by score users on the basis of a score, and the social consequences of that decision. If a score user considers two test certificates to be equivalent for a certain purpose, candidates who attain either one of these certificates will typically gain access to the desired service, position, or status (in this sense, McNamara, 2012 compared language tests to shibboleths). The magnitude of the impact of such decision often depends on the status of the score user: For instance, the decisions made by governments of nation states based on language test scores will typically have a larger impact that decisions made by a local employer.

Validation, as it is conceptualized in this dissertation, thus relies on providing evidence for claims or assumptions made by test developers and score users, which have remained largely unsubstantiated (Kane, 2001, 2012, 2013). The claims and assumptions that require validation range from task selection to score use, and the burden of evidence is shared between test developers and score users, depending on who makes which implicit or explicit claim. Investigating every single assumption or claim that underlies a certain use of a certain score could lead to a never-ending validation process. Kane (2013, 2017) therefore recommends focusing on the assumptions that are potentially the most problematic, or the least likely. Importantly, Kane is rather strict when it comes to evidence for high-stakes score use. The more impact the interpretation or use of a score has on an individual's life, the stronger the evidence should be: All the evidence must support the assumptions made by test developers or score users. Evidence that contradicts an assumption or a claim nullifies the validity argument.

A NOTE ON POLICY

Before introducing the research goals of this dissertation, it is important to briefly introduce the topic of policy analysis, which will be considered more elaborately in Chapter 7. For now it is important to define policy, and to add a disclaimer concerning the research goals and the research design that will be described below.

The term "policy", as it is used in this dissertation, refers to the rules and measures written down in national and institutional laws and regulations to pursue a valued goal or attain a target (Grin, 2003), and to actions taken by key actors in the policy-making process (Ball, Maguire, & Braun, 2012). This definition incorporates two levels of power identified in Wilson (2006): the highest level,

where policy objectives are defined – policy as text (Ball, 2015, p. 2) – and the lower levels, where policy is enacted. It is at this lower level that the success or failure of a policy is often decided (Wilson, 2006). In the context of this dissertation the key actors are people who impact or implement university entrance policy (such as civil servants, deans, educational directors, and university admission officers), and the measures taken to implement a policy are the rules stipulated in the university entrance requirements (see Table 1.2).

Theoretical models (Lasswell's, 1956 model has been particularly influential) show policy as a rather linear sequence of stages, starting with the definition of a problem and ending with a policy evaluation (Jann & Wegrich, 2007). Empirical studies have shown this model to be an idealization, however (Fischer, 2007; Raaper, 2016). Real-world policy making is influenced by budget concerns, ideology, constraints imposed by pre-existing policies, think tanks, and the like (Jann & Wegrich, 2007). Because policy is essentially a patchwork of decisions (Ball, 2015), the policy goals may be ill defined or absent, and the rationale for policy measures may not be specified (Jann & Wegrich, 2007).

Since admission policies of the Flemish universities did not stipulate goals or rationales, assumptions were deduced from the admission requirements. The assumptions formulated in the following sections should thus be seen as the starting point for research, and as a falsifiable basis for the formulation of research goals. This process of teasing out assumptions and arguments from policy texts for the purpose of empirical policy research is not exceptional. Quite the contrary; Fischer (2003) considers it an integral part of the policy analysts' job.

Making assumptions about assumptions may be hazardous, however. As such, policy makers were consulted in order to explore the foundation for the admission regulations. Chapter 7 confirms that most policy did indeed hold assumptions 1, 3 and 4 to be true. Assumption 2 is not for test users, but for test developers to investigate.

RESEARCH GOALS

The research goals identified in this research project concern different aspects of the Flemish university entrance policy. The first four goals focus on empirically examining assumptions that are present in the policy texts, the fifth goal is about determining whether an assumption, held by policy makers, can be empirically verified. Research goal 6 concerns the foundation of the admission policy itself.

Research goals 1 – 4: Examining assumptions

Based on the university entrance requirements and on the assertions made by the STRT and ITNA test developers, four major assumptions (A1 – A4) can be identified on which the university entrance policy towards international students rests. Each assumption is introduced here, and will be examined in detail in this dissertation. These assumptions were originally deduced from the admission requirements, but as Chapter 7 shows, they represent commonly held policy maker beliefs.

Constructs and levels

The first two assumptions concern the adequacy of the target language level (B2) and the language test constructs. They are investigated in Chapter 1 and Chapter 2 of this dissertation.

A1 *B2 is an adequate threshold level to decide on international L2 students' access to a Dutch-medium university in Flanders.*

The first assumption is of a somewhat different order than the subsequent ones, since the B2 demand is central to every university entrance policy. While some universities require international L2 students of literature and linguistics to pass a C1 test (e.g., University of Antwerp), Flemish universities have B2 as a common entrance requirement.

"Adequate threshold level of language proficiency", here, is considered as the point below which the language level of L2 students will likely prevent successful participation in academic studies at university, but above which such participation may be possible. This conceptualization of an entrance level as a minimally acceptable cut off point relies on McNamara & Ryan (2011) and on Carlsen, (2017), who distinguishes between strong and weak interpretation of entrance requirements. The former implies that students who pass the entrance requirements are expected to succeed in the target setting, while the latter denotes that students who do not meet the entrance requirements, are expected to be unsuccessful in the target setting.

If policy makers did not believe that the B2 level was an adequate threshold level for university entrance, they would either knowingly admit students who are likely to struggle with the real-life language demands, or deny entrance to students who meet those demands. Since both options are unlikely and morally dubious, we hypothesized that, like in many other countries (Xi, Bridgeman, & Wendler, 2013), policy makers consider the B2 level an acceptable university entrance level. Some policy texts explicitly state that the B2 level indicates sufficient knowledge of Dutch as a medium of instruction (Universiteit

Antwerpen, 2016, p. 3; VUB, 2014, p. 22). Since no entrance policy differentiates between different skills, we also hypothesized that university admission officers consider B2 an acceptable entrance level for receptive and productive skills.

A2 *STRT and ITNA are representative for the academic language requirements at Flemish universities.*

Following the structure of Figure 1.2, verifying the second assumption is the responsibility of test developers, as long as the test is used for a purpose promoted by the developer (Kane, 2013; but also see Norris, 2008 for the importance of a clearly stated test goal). STRT is marketed as a B2 test designed for people who intend to pursue higher education at a Dutch-medium institution (CNaVT, 2016a). As such, the STRT developers claim that the test can be used to assess the Dutch language proficiency of learners who intent to attend a Dutch-medium university or university college program in Flanders or the Netherlands.

ITNA is used as an achievement test at the end of an L2 learning trajectory offered at the language centers that develop it, but the ITNA website explicitly refers to the test's use as a university entrance test (ITNA, 2016). Actually, more than 70% of the ITNA candidates take the test for the purpose of university admission (Interuniversitair Testing Consortium, 2015), and international university staff members who need to meet the B2 requirement for Dutch are required to pass ITNA as well (Departement Onderwijs en Vorming, 2015). Moreover, the IUTC, the consortium developing ITNA, mentions the first year of higher education as the target language use context (Interuniversitair Testing Consortium, 2015, p. 12).

Since language test developers are responsible for validating their tests in contexts they explicitly promote or knowingly allow (Kane, 2013), and since both STRT and ITNA advertise the use of their tests for the purpose of university entrance, both test developers can be held accountable for validity claims pertaining to task selection, rating, and inference made in this context (Kane, 2013; Weigle & Malone, 2016; Weir, 2005). Both STRT and ITNA promote the use of their tests for more than one purpose. This dissertation focuses on the one purpose they share: admission to university in Flanders.

Selection & Discrimination

Assumption 3 and 4 are concerned with how consistently the university admission policy ensures the admittance of students who have B2 proficiency.

A3 *STRT and ITNA can be considered equivalent measures of B2 Dutch language proficiency.*

Both STRT and ITNA claim a link to the B2 level, and both tests include rating scales that are based on CEFR level descriptors. Likewise, all Flemish universities accept both STRT and ITNA as measures of B2 ability (Universiteit Gent, 2016; KU Leuven, 2016; Universiteit Antwerpen, 2016; Universiteit Hasselt, 2016; Vrije Universiteit Brussel, 2014). Certificates of both tests have the same legal value in the admission process. Empirically investigating Assumption 3 is the focus of Chapter 3 and Chapter 4.

A4 *Students with a Flemish high school degree have obtained Dutch language proficiency at B2 level.*

The Flemish decree regarding the language regulations in higher education is actually more lenient than this assumption. Article II.193 (Vlaamse Regering, 2013) states that everybody with proof of successful completion of any one year in Dutch-medium secondary education should be allowed to register without taking a language test, also if they have not attained the final diploma (they would, in that case, be required to show a different diploma proving that they finished high school somewhere else). In practice this situation quite rarely occurs, except in the case of children of expats.
Assumption 4 is thus more careful than the actual regulations for L2 students are. It is based on the most basic principle of the Flemish university entrance policy, namely that there are no obligatory or binding university entrance tests for students with a degree from a Flemish high school. Chapter 5 examines how safe it is to assume that all these students have achieved the B2 level.

The requirement on which Assumption 4 relies may include other assumptions (e.g., regarding content knowledge), but those are beyond the scope of this dissertation.

In order to investigate to what extent these four assumptions are supported by empirical data, four research goals were identified:

1. Examine the empirical support for the B2 level as an entrance requirement;
2. Compare real-life language requirements at Flemish universities to STRT and ITNA operationalizations;
3. Empirically establish to what extent STRT and ITNA scores can be considered equivalent;
4. Determine whether all students who enter university with a Flemish high school degree pass the B2 threshold.

Research goal 5: Measuring and explaining post-test language gains

Part 1 of this dissertation discusses research conducted to assess the first two assumptions described above. Chapter 1 shows that many professional language testers consider B2 an acceptable entrance level, but expect international L2 students to make language gains while studying. Similarly, the results presented in Chapter 2 shows that students with a B2 level struggle with the real-life linguistic demands at university. Chapters 2 and 7 indicate that policy makers assume that international students' language proficiency level will increase because they live and study in a Dutch-medium context.

If the B2 level is considered an entrance level, and if international L2 students are expected to make language gains over time, it is worth determining whether these gains are actually made. Consequently, the fifth research goal of this study is:

5. Longitudinally track language gains made by international L2 students who have passed STRT, ITNA or both, and explain these gains by analyzing contextual factors.

Research goal 6: Tracing the origins and mechanisms of the Flemish university admission policy

Empirical research conclusions are essential for useful policy evaluations, but they are not the be-all and end-all. Science typically uses a different paradigm than policy making. Academic research is premised on four basic principles: truth, autonomy, independent funding, and peer-driven quality assessment (Wollmann, 2007). These principles do not carry the same weight the highly entangled and politicized domain of real-world policy making, however.

For that reason, Chapter 7 gives voice to policy makers, and shows how the current policy came to be. By highlighting the role of empirical data in the Flemish university admission process and by uncovering the mechanisms that impact the policy-making process, this chapter offers essential information for making realistic policy recommendations.

6. Uncover the mechanisms and assumptions that have shaped the Flemish university admission policy.

AVAILABLE EVIDENCE

Not each of the above-mentioned research goals has remained completely unexplored to date. Some have been the topic of research, but the results are scattered and incomplete. This section provides an overview of the research that has been conducted to investigate the five research goals.

Research goal 1

The first research goal concerns the B2 requirement for international L2 students, which is central to all university entrance policy documents and is reiterated in the STRT and ITNA validity arguments. Nevertheless, there is no publicly available evidence to support the near-universal B2 requirement. Policy texts and legal documents typically refer to the B2 level as "sufficient" or "required", but do not offer any substantiation or proof (e.g., *Besluit van de Vlaamse Regering tot codificatie van de decretale bepalingen betreffende het hoger onderwijs*, 2013; Departement Onderwijs en Vorming, 2015).

Research goal 2

Both STRT and ITNA have produced validity arguments in order to obtain the seal of quality awarded by the Association of Language testers in Europe (ALTE). This Q Mark is awarded to tests that have been successfully audited. An ALTE audit procedure involves a review based on seventeen minimum standards, including the construct, the purpose, the rating procedure, the CEFR link, and the robustness of the statistical analyses. The audit reports themselves are confidential, but both STRT and ITNA have given insight into their preparatory documents (CNaVT, 2014; Interuniversitair Testing Consortium, 2015). These reports and the audit outcomes show that both tests have satisfactory rating procedures (e.g., STRT K = .8; ITNA K = .73) and internal reliability (e.g., in both tests Cronbach's alpha < .9). Both tests employ Rasch modeling to monitor and control the difficulty of the exam. Both test developers claim that candidates who are awarded certification have the B2 level. Internal documents provided by both tests reveal that both ITNA and STRT have been linked to the B2 level of the CEFR using the stepwise approach proposed in Figueras et al. (2005). Validity evidence pertaining to STRT has been presented in peer-reviewed journals (Deygers & Van Gorp, 2015) and books (Deygers, Van Gorp, Luyten, & Joos, 2013), and at conferences (Deygers, De Wachter, Van Gorp, & Joos, 2013). Research concerning ITNA has been presented at conferences (e.g., De Geest, Steemans, & Verguts, 2015; Steemans & Vlasselaers, 2013).

Even though STRT and ITNA have provided quantitative data concerning internal reliability, rating, and the like, the evidence regarding

research goal 1 – representativeness for the target language use context – is scattered. STRT is specifically developed for future students at a Dutch-medium institution of higher education, and a recent STRT publication contains the claim that students who have passed the test might well function adequately in the target setting (Maes, 2016). The current operationalization of STRT (see Appendix 1) is based on a needs analysis, on the CEFR descriptors for the B2 level, and on a literature review focusing on language requirements in the academic domain (CNaVT, 2014). The needs analysis (Gysen & Avermaet, 2005; Van Avermaet & Gysen, 2006), conducted in 2000, applied factor analysis to questionnaire data, to identify the perceived needs of Dutch language learners (N = 700) and Dutch language teachers (N = 800). The primary needs in the educational domain were identified as taking a written and an oral exam, attending class, and writing a paper (Gysen & Avermaet, 2005, p. 55).

ITNA makes no claims regarding future performance in the target setting, but the developers do state that most candidates take the test for university entrance purposes. Consequently, the ITNA test tasks have been designed to meet the needs of this population (ITNA, 2016, p. 11). Additionally, the computer test is claimed to reliably and validly indicate whether a candidate has achieved the B2 level (ITNA, 2016). Nevertheless, the rationale provided to support the use of the test tasks (see Appendix 2) does not refer to empirical research, needs analyses or target context research conducted by the test developer. The validity argument for the task types relies mainly on the B2 descriptors themselves (e.g., the reading structure task), on test developer experience (e.g., the word transformation tasks), or on published research conducted in a different context (e.g., cloze: Bachman, 1985; dictation task: Cai, 2013). No explicit rationale is given for the oral argumentation task or the presentation task.

Research goal 3

At all Flemish universities STRT and ITNA are accepted as equivalent measures of B2 proficiency. Nevertheless, test users who consider them legally or linguistically equivalent have provided no evidence to back the implicit assumption that the B2 ability measured by both tests is comparable. However, a pilot study that was conducted jointly by the STRT and ITNA test developers did investigate test equivalence. The results show strong correlations (see Cohen, 1988) for the written (r = .77, p <.000, N = 77) and low correlations for the oral components (r = .15, ns, N = 38) (Deygers & Luyten, 2012; Van Gorp, Luyten, De Wachter, & Steemans, 2014).

Research goal 4

The assumption that students who have obtained a Flemish secondary education degree will also meet the B2 language requirements, has not been confirmed in publicly available research. One study (Van Houtven & Peters, 2010; Van Houtven, Peters, & El Morabit, 2010), conducted at four Flemish university colleges (N = 176), showed that not all first-year students passed a summary task of the PTHO, the B2 test which preceded STRT. Secondly, De Wachter et al. (2013) showed that there is substantial variation in the L1 proficiency of first-year students at the University of Leuven. However, the extent to which first-year Flemish university students meet the B2 demands has not been investigated.

There are no centralized tests at the end of Flemish secondary education, but there are common attainment targets or educational goals, which are operationalized and tested by teams of teachers or by individual teachers. These official attainment targets have not been linked to the CEFR. In programs that typically prepare for university, educational goals are comparable to B2 tasks. In some vocational programs, however, the level of the attainment targets may be below B2 (e.g., Onderwijs Vlaanderen, 2015).

Research goal 5

There is very little research into language gains made by international L2 students who do not take additional language courses. To date, research in that field has primarily focused on measuring writing gains in English-medium contexts (Knoch, Rouhshad, Oon, & Storch, 2015; Storch, 2009). The available results suggest that, international L2 students who do not receive additional language classes are unlikely to make major language gains. Prior to this dissertation, no research of this kind had been conducted in a Flemish setting.

Research goal 6

To date, the mechanisms of policy making at Flemish universities has not been the topic of research. Internationally too, no policy research has been conducted in the specific domain of university admission.

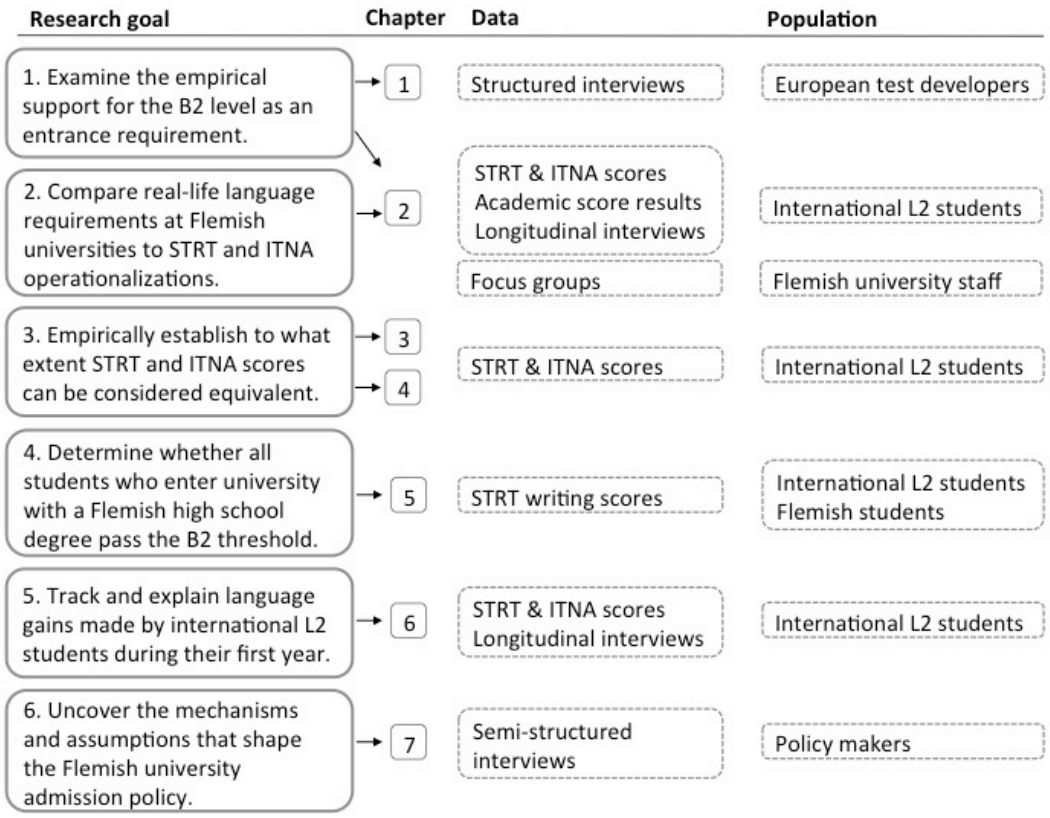
RESEARCH DESIGN

Gathering and analyzing data concerning language gains, test scores, experiences and opinions requires a varied set of approaches within the same overall design. Mixed methods research uses both quantitative and qualitative approaches to tackle the same overarching research goal in one research project (Davies, 2010;

Xi, 2010). Since every chapter includes a dedicated research methodology section, this introduction to the research design primarily serves to show how the studies are connected, rather than to offer detailed information on the approaches used to investigate every research goal. Figure 1.3 summarizes which chapters focus on which research goals, using which data and which research population.

This dissertation includes data collected among different research populations. The populations, and the data obtained from them, are introduced below, but more detailed information concerning data and data analysis, is given when relevant to specific chapters.

Figure 1.3. Research design



Research populations

European test developers

N: 30
Data: Structured interviews
Period: November 2014 – April 2015
Analysis: Qualitative (Chapter 1)

In order to determine common characteristics in the myriad of university entrance policies across Europe, 30 experts representing 28 European regions were interviewed. All respondents were professionally involved in language

testing, and all were members of ALTE (Association of Language Testers in Europe). The outcomes of these interviews are discussed in Chapter 1.

Flemish university staff

N:	24
Data:	Focus group
Period:	January and February 2014
Analysis:	Qualitative (Chapter 2)
Appendix:	5

In January and February 2014, 24 university staff members from the two largest Flemish universities (Ghent University and KU Leuven) took part in six different focus groups in order to delineate the minimally acceptable university entrance language level and to establish which task types can be considered essential for students to master upon university entrance. Chapter 2 shows the results of these focus groups.

International L2 students: $L2_P$ (Pilot)

N:	11
Data:	Semi-structured interviews
Period:	October – December 2012
Analysis:	Qualitative (Chapter 2)
Appendix:	3

These participants were international L2 students who had enrolled at Ghent University after passing STRT or ITNA. They were interviewed at the start and at the end of their first semester at Ghent University, as part of a pilot study. The analyses of these interviews – included in Chapter 2 – offered information concerning the real-life language demands at Flemish universities.

International L2 students: $L2_F$ (took STRT in Flanders)

N:	135
Data:	Semi-structured interviews
	STRT & ITNA scores
	STRT test/retest scores
	Academic score transcripts
Period:	June 2014 – May 2015
Analysis:	Mixed Methods (Chapters 2 and 6)
	Quantitative (Chapters 3, 4, 5)
Appendix:	4

During the summer of 2014, 135 L2 International L2 students who planned to enroll at a major Flemish university (Ghent University, KU Leuven, University of Antwerp) took STRT and ITNA within the same week. These respondents had registered for ITNA and had agreed to also take STRT free of charge. Their test scores were used in analyses in all chapters of this dissertation, except for Chapters 1 and 7. Twenty respondents within this population agreed to be part of a longitudinal study that traced the linguistic, academic and social hurdles they encountered during their first year at university. The longitudinal data gathered from these twenty respondents, are discussed in Chapters 2 and 6 (which also includes STRT retest data).

International L2 students: $L2_I$ (took STRT at their home university)

N:	526
Data:	STRT test scores
Period:	May 2015
Analysis:	Quantitative (Chapter 5)

The scores of STRT candidates who had studied Dutch at their home university and taken the test there were used to compare the pass probability and score profiles of international L2 students to Flemish students.

Flemish Students

N:	159
Data:	STRT writing scores
Period:	October 2015
Analysis:	Quantitative (Chapter 5)

In order to determine whether incoming students with a Flemish high school degree have B2 language proficiency, 159 first-year students of business studies with a Flemish high school background took two STRT writing tasks. In Chapter 5 their scores are compared to the $L2_F$ and $L2_I$ STRT writing scores.

Flemish Policy makers

N:	15
Data:	Semi-structured interviews
Period:	December 2016 – February 2017
Analysis:	Qualitative (Chapter 7)

The civil servants responsible for the university entrance language regulations at governmental level and the policy makers responsible for the admission criteria at the five Flemish universities were recruited in order to examine how the

Flemish university admission policy takes shape. The results of this study are examined in Chapter 7.

Methodology

All qualitative data, and all qualitative data analyses were double-checked. Every interview was audio recorded and transcribed by trained transcribers. All transcripts were checked by the researcher, and were partly double coded. All coding was done using NVivo 11 For Mac, except the double coding for the interviews with European test developers, which was manual. Table 1.3 shows that the level of inter-coder agreement was consistently high.

Table 1.3. Double coding inter-rater agreement

	Double coded	Agreement
European test developer interviews	30%	86%
Flemish university staff	20%	90%
$L2_F$ interviews	20%	90%

In order to check the accuracy of the transcribers, three randomly selected interviews were fully transcribed. Prior to calculating the similarity of the samples, the transcripts were preprocessed. First, transcripts of interviewer speech and participant speech were separated, since it was assumed that accent familiarity might impact the overlap in the transcripts. The overlap between the transcription pairs was checked using the two indices used in information retrieval methodology. Both the Jaccard index *J* and the Sørensen–Dice index *QS* are commonly used and accepted indicators of similarity between two samples (Gomaa & Fahmy, 2013). In both cases, two identical datasets yield an index of 1 and two utterly dissimilar sets would result in an index of 0. The results show a very large degree of overlap between the two sets of interviewer (J = .9; QS = .95) and interviewee transcripts (J = .77; QS = .87).

All quantitative data – test scores, language gain indices – were analyzed using *R* (descriptive and inferential statistics), Facets (Multi-Faceted Rasch analysis), and Python (text pre-processing and text similarity indices). Within the *R* environment, the following packages were used: car (simple and multiple regression), *exacti* (McNemar's test), *ggplot2* (data plotting), *Hmisc* (correlation), *irr* (inter-rater reliability), *MASS* (principal component analysis), *Mlogit* (multinomial regression), *Pastecs* (data plotting), *Pgirmess* (Kruskal-Wallis test), *Plyr* (summary statistics), *prob* (pass probability), psych (descriptive statistics, examining normality), *QuantPsyc* (regression). The specific methods used to analyze the data will be explained throughout the dissertation, when relevant to the research goal at hand.

STRUCTURE AND RELEVANCE OF THIS DISSERTATION

This dissertation addresses several gaps in the existing literature. It examines the impact of policy measures, and determines whether these measures have the desired effect. Policy impact research of this kind has been performed in the context of public (Nagel, 2002), or environmental policy (EEA, 2001), and even in the context of K12 school policy (Ball, Maguire, & Braun, 2012; Grin, 2000, 2003), but university entrance policies have typically not been the subject of large-scale, triangulated studies (McNamara & Ryan, 2011). Similarly, the idea of justice, as related to high-stakes language testing for admission purposes, has remained largely underexplored (Khan & McNamara, 2017), and there is no clear definition of what a just admission testing policy might look like. This dissertation offers a way forward regarding both issues.

The chapters following the introduction are based on research papers which have been published or accepted by international peer-reviewed journals, and which have been edited to fit the format of this dissertation and to avoid redundancy. Each chapter focuses on one aspect of the Flemish university entrance policy, but also has wider implications for the language testing or applied linguistics community.

Part one: Constructs & levels

The first chapter is based on structured interviews with 30 expert respondents. The results show that the CEFR is omnipresent in European university entrance language tests and that the B2 is the most commonly used level in that context, but often without empirical support. Prior to this study, no research had either examined the communalities between different European policies, or the impact of the CEFR on these policies. Chapter 1 frames the dissertation in a wider context. As of chapter 2 the focus narrows to Flemish policy.

Chapter 2 brings together the opinions and experiences of 24 university staff members and 31 international L2 students ($L2_P$ and $L2_F$). The outcomes of this study reveal that the actual receptive language requirements of university studies most likely exceed the expected B2 level, and that the Flemish entrance tests include language tasks that are of little importance in the real-life university-based studies of first-year students. By having a group of candidates take both STRT and ITNA, it was possible to track the progress in the target context of people who had actually failed one of the two entrance tests. This is quite possibly the first study to partially bypass the truncated sample problem (Alderson, Clapham, & Wall, 1995) by using this approach. Drawing on the observation that a number of students who failed the entrance test actually managed quite well at university, this chapter also discusses matters of justice related to admission testing.

Part two: Performance and equivalence

The third chapter examines the implicit claim that STRT and ITNA can be considered equivalent measures of the same level of language proficiency. Additionally, this chapter aims to explain not only the strength but also the nature of the relationship between the two tests. To the best of my knowledge, no studies have yet been published in which researchers had access to detailed rater data from two different, high-stakes entrance tests in order to compare level and construct equivalence. The results discussed in Chapter 3 reveal important discrepancies between STRT and ITNA.

The following chapter focuses on the equivalence of corresponding rating criteria. STRT and ITNA use the same criteria to assess performances on the oral argumentation tasks and presentation tasks. The level descriptors used to score these criteria are based on the same CEFR descriptors. The results show, however, that the scores – assigned to the same candidates performing nearly identical tasks – deviate significantly. Prior to this research no study had yet assessed the equivalence of CEFR-based rating descriptors across different tests.

Chapter 5 investigates whether all Flemish candidates have a B2-level in Dutch upon university entrance, and whether L1 test takers outperform L2 candidates who learned Dutch at home or in Flanders. The results show that, even though the Flemish group outperformed both groups of L2 candidates, not all Flemish candidates reached the B2 level. To date, no study had compared L1 and L2 performance on a centralized high-stakes B2 university entrance test.

Part three: Gains & context

The sixth chapter tells the story of twenty international L2 students ($L2_F$) during their first year at three universities in Flanders, Belgium. The results show that after eight months at university, the respondents had only made progress in terms of written fluency. No other gains in the oral or written modality were observed. The qualitative data used to explain the language gains show that the respondents experienced little institutional support, and that few respondents had gained access to the L1 academic community. Longitudinal studies that focused explicitly on language gains by international L2 students who received no additional language support, are very limited in number, and have so far only considered writing gains. Within this context, chapter six presents the first study to measure speaking gains. No existing studies have considered personal and academic experiences of L2 learners to explain oral and written language gains.

Part four: policy, conclusion & implications

The final component of this dissertation includes a discussion of how the university admission policy is made. Chapter 6 shows how stakeholders' interests have impacted regulations and shows that empirical data have not had a demonstrable influence on the policy measures in place. These observations are connected to the conclusions drawn from the empirical research to formulate implications and recommendations based on Fischer's model of policy evaluation in the final Chapter. This triangulated and multifaceted approach to policy evaluation is new to the field of language assessment.

A NOTE ON COMPOSITION

The first six chapters of this dissertation are based on research papers that have been submitted to, accepted or published by peer reviewed journals or books. In order to avoid unnecessary repetition across chapters, certain sections of the original papers have been omitted from the chapters and are included in this introduction. The sections in this introduction dealing with the context of research, with the approach to validation, or with the research population have been adapted from these research papers (Deygers, 2017; Deygers, Van den Branden, & Peters, 2017; Deygers, Van den Branden, & Van Gorp, 2017; Deygers, Van Gorp, & Demeester, 2017).

PART 1
LEVELS & CONSTRUCTS

> The first part of this dissertation investigates two aspects of the Flemish university entrance policy: required language level, and the representativeness of the main gatekeeping tests for the target context.

In order to situate the Flemish university entrance policy in a wider context, the first chapter offers an overview of the language requirements for international L2 students throughout Europe. The data show that in many European countries international L2 students are required to prove language proficiency at the B2 level. At the same time, however, the B2 requirement seems to have become rather pervasive without much solid empirical evidence. In most European contexts surveyed – Flanders included – there was little or no publicly available empirical evidence to support the required university entrance level.

In the second chapter the focus shifts to Flanders. Chapter two shows that the B2 requirement, especially for receptive skills, does not correspond to the actual language demands at university. Additionally, the data indicate that the two major university entrance tests used in Flanders, STRT and ITNA, may operationalize or prioritize language tasks that are not necessarily important in the target setting, or are not yet expected of incoming students. Drawing on the observation that L2 students are assessed on the basis of tasks that their L1 peers are not expected to perform, this chapter also reflects on matters of justice in university entrance testing.

Chapters 1 and 2 are based on:

> Deygers, B., Zeidler, D., Vilcu, D., & Carlsen C.H. (2017, in press). One framework to unite them all? The use of the CEFR in European university entrance policies. *Language Assessment Quarterly*.

> Deygers, B., Van den Branden, K., Van Gorp, K. (2017, in press). University entrance language tests: a matter of justice. *Language Testing*.

The papers have been formatted to fit the structure of this book. Certain components (e.g., *methodology, participants*) have been rewritten to avoid redundancy and benefit readability.

CHAPTER 1
UNIVERSITY ADMISSION POLICIES ACROSS EUROPE

The aim of this chapter is to investigate how widespread the use of the CEFR, and more specifically the B2 level, is in European university entrance testing. As such, Chapter 1 frames the Flemish policy in a larger context, and shows that the lack of publicly available evidence to support an admission policy is not unique to Flanders.

The goal of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) is to promote the free movement of people and ideas by increasing the transparency across educational systems through the common use of the same proficiency levels (Van Ek, 1975). The first drafts of what was to become the CEFR appeared in the early 1970s with the development of the *Threshold* level, later called B1 (Van Ek, 1975). Through the years new levels were added (Van Ek & Trim, 1991b; Van Ek & Trim, 2001), existing levels were refined (Van Ek & Trim, 1991a), and some thirty years after the first drafts, the project culminated in the "blue book" we know today. The CEFR proposes six consecutive levels of language proficiency, ranging from A1 through to C2. It focuses on what learners putatively can do with language and includes 53 illustrative scales, which list language-independent descriptions of each proficiency level for a given skill or ability. Arguably, these illustrative scales have become the most influential, but also the most heavily criticized aspect of the CEFR (Little, 2007; Figueras, 2012).

CEFR criticism is usually either political in nature or content-related (Figueras, 2012). The politically oriented criticism sees the CEFR as an instrument of power that encourages a simplistic, level-driven logic at the expense of a needs-based, user-driven policy (McNamara, 2007; Shohamy, 2011). According to this strand of criticism the CEFR levels are often used as a normative standard that provides policy makers with an easy tool to assist in gatekeeping (Fulcher, 2004; 2012). In a defense against this criticism, North warns against confusing intended CEFR use with actual but inappropriate use (North, 2014a), and considers any normative use of the CEFR to fall under this category (Martyniuk, 2010; North, 2014a). Furthermore, so the argumentation goes, the CEFR discourages a simplistic level-based gatekeeping policy, because it stimulates discussion between decision makers and language experts (Porto, 2012). Content-related criticism on the other hand focuses on those aspects of the CEFR that are problematic in the context of language testing – the field where the CEFR's influence has been most keenly felt (Little, 2007). Some authors tackle the foundation of the CEFR, pointing out that it lacks empirical validation (Fulcher, 2012b), ignores insights from second language acquisition theory (Alderson, 2007;

Little, 2007; Hulstijn, 2007), does not pay equal attention to all skills (Weir, 2005b; Alderson et al., 2006; Staehr, 2008) or does not cover the full range of levels in every scale (Alderson et al., 2006). It is now commonly recognized that the theoretical support for the descriptors of the receptive skills is not quite robust (Alderson, 2007; North, 2014a), but a growing body of research has been providing language-specific empirical validation for the CEFR (Salamoura & Saville, 2009; Carlsen, 2014). A second strand of criticism related to the content focuses on the deficiencies of the CEFR as a common reference point in language testing. This line of criticism maintains that the levels are not equidistant as they contain overlaps and inconsistencies, and that the descriptors are general, language-independent, containing impressionistic terminology (Alderson et al., 2006; Alderson, 2007; Papageorgiou, 2010, Fulcher, 2012).

This fierce criticism towards the CEFR may seem somewhat overstated in comparison with the document's intended use, which appears relatively modest. The authors have repeatedly pointed out that the CEFR descriptors were meant to be general and that the levels were never meant to make up an interval scale, so in part the content-related criticism points out a characteristic that was purposefully built in to the CEFR (Little, 2007; North, 2014a). In their defense, CEFR authors also state that in spite of the general nature of the CEFR descriptors, professionals in organizations such as ALTE and EALTA (European Association for Language Testing and Assessment) interpret the CEFR levels in a uniform fashion (North, 2014a). When it comes to the use of the CEFR, its authors insist that it was not meant as a normative document, but as a malleable heuristic that stimulates reflection, facilitates discussion about language learning, and aids curriculum planning and language certification (North, 2007; 2014a; 2014b). The actual use of the CEFR has not always been in tune with its intended use however (North, 2014a), and many language tests across Europe have experienced pressure from stakeholders to link their scoring system to the CEFR (Fulcher, 2004), or have been redeveloped with the CEFR descriptors in mind (Galaczi, ffrench, Hubbard, & Green, 2011). Also, as many countries and institutions in the EU now use CEFR levels to set legally binding citizenship requirements (Van Avermaet & Pulinx, 2013), curriculum goals (University of Cambridge, ESOL Examinations, 2011) and university entrance demands (Xi, Bridgeman, and Wendler, 2014), language testers in Europe have been required to follow suit.

Fifteen years after its publication, the CEFR has inspired a large body of literature, and has gained critics and champions. It has been praised for facilitating dialogue (North, 2014a) and denounced for causing validity chaos (Fulcher, 2012b). It has been studied in the context of rating scale design (Harsch & Martin, 2012), and it has been the subject of sociolinguistic debates (Roever & McNamara, 2006). It has been adopted by users in other contexts and on other continents (Negishi, Takada, & Tono, 2013), and has been accused of linguistic

imperialism for it (Curtis, 2015). The CEFR has led to a lot of language testing research, but its actual impact on the lives of test takers has remained largely unexplored. Still, because of its effect on language tests, it has potentially affected the lives of millions. The study presented in this chapter focuses on the CEFR's impact in one specific field: university entrance language testing for L2 students from abroad.

RESEARCH QUESTIONS

There is little reliable information about the impact of the CEFR on L2 university entrance language policy across Europe, and the information that is available is fragmented or scattered across the websites of Europe's universities and departments of education. As such, the existing literature does not offer the opportunity to frame the Flemish situation in a wider context. For that reason, an explorative study was set up, with two guiding questions:

RQ1: What are the common features of university entrance language requirements for international L2 students across Europe?

RQ2: What is the impact of the CEFR on university admission policies and university admission language tests in Europe?

These research questions help to frame the Flemish university entrance language policy. The results offer information concerning the empirical ground for using the B2 level as a threshold for university admission. As such, the data presented in this chapter are relevant to Assumption 1 (B2 is an adequate threshold level to decide on international L2 students' access to a Dutch-medium university in Flanders).

PARTICIPANTS & METHODOLOGY

Respondents

The information required for this study is only known by relatively few people who work in the context of university admission or university entrance language testing. Consequently, collecting data via randomized sampling was of little use and respondents were chosen through purposeful selection (Freeman, 2000), which implies identifying knowledgeable and information-rich respondents (Reybold, Lammert, & Stribling, 2013). All respondents were professionally involved in language testing, and in many cases in the development of language

tests for university entrance. The researchers contacted members of European language testing organizations that are full members or affiliates of the Association of Language Testers in Europe (ALTE). Thirty-nine organizations were asked to take part in the study, and in the end representatives of 30 organizations operating within 28 states or regions with autonomy over educational matters agreed to participate. Although test operationalization was not a selection criterion, all university entrance language tests included in this study (N = 25) are skill-based, seven of which also feature a grammar and vocabulary section.

Scope

For reasons of readability we will use the term "context" throughout the text when referring to both nation states and regions that have (quasi) autonomy over educational matters. Consequently, even though this study covers 26 nation states, this chapter reports on 28 contexts.

Table 2.1. Countries and regions surveyed

R1	Austria	R17	Luxembourg
R2	Belgium (Flanders)	R18	Malta
R3	Belgium (Wallonia)	R19	Netherlands
R4	Bulgaria	R20	Norway
R5	Czech Republic	R21	Poland
R6	Denmark	R22	Portugal
R7	Estonia	R23	Romania
R8	Finland	R24	Slovenia
R9	France	R25	Spain (Basque)
R10	Germany	R26	Spain
R11	Germany	R27	Sweden
R12	Greece	R28	Switzerland
R13	Hungary	R29	United Kingdom
R14	Ireland	R30	United Kingdom
R15	Italy		
R16	Lithuania		

Because of the independent educational policies in Wallonia, Flanders (both in Belgium) and the Basque region (Spain), these regions were considered distinct contexts. In other countries, such as Switzerland, there are different linguistic regions, but the policy and the entrance tests in these regions share the same vision and characteristics, and as such they are counted as one context. In some cases there were two respondents per context because there was more than one full member of ALTE for that country, and more than one member wished to

participate (R10 and R11, R29 and R30). These doubles allowed the researchers to compare the answers to the factual questions. In both cases the same information was given by both respondents. Table 2.1 lists the codes of the respondents and the countries surveyed in this study.

Data collection and analysis

Since it was considered important to collect data that shed light on each individual context, but also allowed for meaningful comparison, structured interviews were used. Respondents were asked a fixed set of questions, but they were free to contextualize their answers in a way that questionnaires cannot accommodate (Schwartz, Knäuper, Oyersman, & Stich, 2008). The interview scenario consisted of three parts: (1) one concerning the university entrance policy, (2) one concerning the entrance tests, and (3) one concerning the interviewee's personal opinions about the CEFR and university entrance language tests (not included in this chapter, see Deygers, Zeidler, Vilcu, & Carlsen, 2017). Respondents received the factual questions three days beforehand via e-mail and were encouraged to look up any information they did not have on hand. The scenario was trialed twice in interview conditions, and the interviews were conducted by four trained researchers via Skype and were recorded using the Audacity software.

All interviews were transcribed verbatim. The transcriptions were checked by two researchers who replayed the interview, corrected any inaccuracies, coded the transcripts independently, and compared codes and outcomes afterwards. Coding was done both manually and using the qualitative software NVivo For Mac, since a combination of manual and computer-assisted coding is likely to yield the most reliable results (Welsh, 2002). Both researchers used a set of agreed-upon a priori codes that corresponded with the questions asked in the structured scenario. After the first round of coding, the exact inter-rater agreement was checked for the a priori coding categories (86.4%), based on a random sample representing 30% of the total transcribed text. Both coders also employed a grounded theory approach (Glaser & Strauss, 1967; Miles & Huberman, 1994), which means that they also coded salient issues that emerged from the data. This double approach allowed the researchers to compare factual data across contexts, but also to spot opinions or views that surfaced during the interviews without being explicitly probed for. After coding independently, the researchers discussed their coding categories over several meetings until consensus was reached. The complexity of the coded data was reduced by quantifying recurring practices and patterns (Ziegler & Kang, 2016). These quantifications are presented in the tables below.

All interviews were conducted in English and the quotes included in this article are literal transcriptions that have been lightly edited for the sake of readability. Editing was restricted to correcting grammatical flaws, and omitting word repetitions and filled pauses. Throughout the chapter, "I" will be used to refer to the interviewer, and respondents will be referred to by their code (i.e., R1). Misinterpretation of what was said during the interview is always possible. For that reason, the transcriptions were sent back to the respondents so they could comment on any factual flaws. When this happened, the information was used in the data analysis, but the original transcripts were not altered. As a final step in the verification of the interpretation of the data, the respondents, as well as the members of ALTE's CEFR special interest group received a prefinal draft of this chapter, which they could amend. No respondent requested any revisions.

RESULTS

In 23 contexts, passing a language test is mandatory for university entrance for L2 students. In three of these countries, only one centralized test is accepted for first-year university entrance, but the twenty other contexts all have a system where multiple tests are used for the same high-stakes purpose. In thirteen contexts both centralized tests and local tests – developed by the receiving university itself – are accepted (see Table 2.2).

Table 2.2. Tests accepted for university entry (N = 28)

	Number of contexts
Multiple centralized and local tests	13
Multiple centralized tests	5
One centralized test	3
Multiple local tests	2
No test	5

The respondents did not generally regard local tests positively and express concern over their quality, transparency and comparability. Out of the fourteen respondents from contexts where both centralized and local tests can grant university access, eight respondents wished to streamline the university entrance system.

R10 [The locally developed tests] don't apply piloting, pretesting and statistical analysis, so it's not astonishing that the results were very heterogeneous and the exams are of a different level of difficulty.

In 22 contexts there is a university entrance language requirement that is expressed in CEFR terms (see Table 2.3). In five contexts (Finland, Luxemburg, Portugal and Spain, including the Basque region) there are no specific language requirements, and in Sweden the requirements are not CEFR-related.

Table 2.3. CEFR level required for university entrance (N = 28)

	Number of contexts
A2 and B2†	1
B1 or higher †	1
B2	9
B2+	1
B2 or C1 †	8
C1 or higher †	2
Requirement not CEFR-related	1
No requirement	5

Note. † varies by program

The findings of the study confirm the assumption that B2 is the most commonly required level for university entrance across Europe: in nine contexts B2 is the only required level, and in another ten it is one of the required levels. But even though B2 is the most commonly used CEFR level for university entrance, there is no agreement among the 30 individual respondents that B2 users have the linguistic resources required to function at the start of academic studies at university (see Table 2.4).

Table 2.4. Is B2 enough to function linguistically at the start of university? (N = 30)

		Number of contexts
No		7
	B2+ as a minimum	(4)
	C1 as a minimum	(3)
Not Quite		8
	If additional language support is offered	(3)
	Depends on student needs	(5)
Yes		10
	B2 is enough	(3)
	B2 is the absolute minimum	(7)
Don't know		5

The respondents may doubt the B2 level as an adequate level for university entrance, but test developers are rarely the ones who make the decision on where to set the entrance level requirements. In sixteen contexts (see Table 2.5) the

decision on language level requirements for university entrance is partly or completely up to university or faculty staff. In four out of sixteen instances the ministry decides on a general rule, but the university determines the actual level, and in the twelve remaining cases the level requirements are left to the university entirely. In seven contexts there is a national regulation stipulating the level requirements.

Table 2.5. Who decides on language requirements (N = 28)

	Number of contexts
No requirement	5
Government (General rule) & university (specifics)	4
Government or ministry of education	7
University, faculty or department	12

In most contexts where a language level is required for university entrance, that level was not determined on the basis of empirical data or needs analyses, the respondents report.

R4 There isn't any official study that said B2 is the right level for students. It's just intuitive, from the practice in universities. Knowing that other countries require B2 they just decided to introduce B2.

R24 [The university] asked us about our opinion, and we said, in Europe it's mostly B2. And they said, okay, it should be B2.

R21 I don't think it's based on any empirical ground. It's just an administrative decision […] Those in charge of making decisions start with the levels. Then the whole field is trying to adjust to the decisions rather than starting with an analysis of the needs […] When the process has been reversed, it is very difficult to turn it around.

In the 23 contexts where L2 university entrance is regulated by language tests linked to the CEFR, language requirements rarely seem to be based on empirical data (see Table 2.6). Only one respondent claimed that the requirements in his/her context were empirically founded.

Table 2.6. Empirical research to support required language level (N = 23)

	Number of contexts	
No Empirical foundation		19
Following other institutions	(6)	
Literature study	(6)	
Unknown	(6)	
CEFR	(3)	
Partly empirically founded		3
Needs analysis	(2)	
Expert counsel	(1)	
Empirically founded		1

This does not imply that the tests themselves are not based on empirical studies. Some tests are based on extensive research and needs analysis, but universities are free to decide on the required entrance level themselves. In many cases universities do not seem to base their requirements on needs analyses, but on common practice or on what their competitors do.

R30 There is empirical research and there was research done when the test was first being designed and when the test was being trialed. [...] Obviously, the official advice is that [people from the accepting institution] should sit down and evaluate the courses; say for a course of that nature, we need a minimum standard of whatever. Now, some universities have done that kind of thing, but others are essentially responding to the market in terms of seeing what other people are asking for. [...] In the sense of actually doing that formal sit down, standard setting session, I think only a limited number have done that. [...] I would be happier if I believed that universities were really doing a standard setting, rather than simply responding to what their competitors and peers were doing.

In quite a few contexts, CEFR requirements depend not on needs analyses, but on financial considerations. Some universities require a CEFR level that is lower than what is required in other universities in order to attract students (R30 above, and R13, R15 below). In at least eight contexts surveyed it is common practice for universities or ministries to use CEFR levels to control the flow of incoming students for financial reasons, to attract a certain type of students, or to control access to the labor market. Level requirements would rise and fall, not because the language needed to participate in academic life changes, but because certain faculties require more or fewer students.

R15 In some universities in the south part where the students are very few, they provide very easy entrance tests [...] while in the north part we have more difficult tests.

R13 Universities can decide to lower the level [...] if they desperately need students.

R4 The situation for international students is different depending on which country they come from. If they are from the European Union they are not obliged to have B2.

R6 The C1 level [...] that's the normal thing. But we now have a special case: medical doctors and nurses in health care that were educated in their homeland. [...] They informed us from the ministry, that they needed people within the health system and they found out that the C1 level, it was too difficult to pass.

R9 We have an institute [of higher education] working with oil and there we don't ask any level, because we need people working with oil.

Of the 30 respondents involved in this study, 27 developed CEFR-related language tests. Most of these respondents stated that the CEFR was used to set or define the level (20), and/or to draw up rating scales (4) and/or to design task specifications (4). Two respondents state that their tests were fully based on the CEFR, and for most respondents, the CEFR is part of their daily practice:

R2 It's always in the background.
R15 Every day I consider the CEFR.
R28 Everything we do is based on the levels.
R30 We refer to it all the time.

The respondents mention three main effects of the CEFR on language testing in Europe. First of all, they feel that the CEFR has brought standardization where there was disharmony. Secondly, the CEFR has promoted skill-based language testing. Thirdly, it has led language tests to adopt a new level structure, which is now so well established that test developers may experience pressure to align with it. Seventeen respondents have experienced an external pressure to align with the CEFR, either politically or economically.

R2 It was in fact the demand [of the funding body] to take into account the CEFR. So because of this, we took the B2 level as a starting point for developing our exam.

R10 The CEFR levels are not precise enough, but if you want to sell an examination or a textbook without indicating the levels, then it will be very difficult on the market, so this is why most of the institutions use the CEFR levels when they want to sell something.

R18 To pass our B2 exam you need to be quite good, but there are other exams which aren't as complete or as strict in their marking and then you get a student who is at a lower level but still gets a B2 certificate and would not manage to pass our exam. So, market-wise, it's creating a bit of a problem here. […] The product has to be marketable, you know?

There are established procedures for aligning tests with the CEFR, but not all test developers have linked their tests to the CEFR using a robust procedure. Respondents of eleven contexts spontaneously mentioned examples of unfounded CEFR links in university entrance tests used in their context. In some cases there was no procedure at all.

R12 We had tests before, I have all the presidential decrees here, and now the announcement is that it is CEFR compatible in all levels […] So these four levels became six now, according to the CEFR framework.

I Where does [normative CEFR use] come from you think?

R30 Cause it's easy I suspect. "Ooh it's there, it's in the book. Yes! B2 is very good, ooh that's fine, yes!" Boom, job done. I think there's a lot of that.

R18 It's not the CEFR descriptors which are creating a problem, really. It's those who are interpreting. […] Everybody else is issuing certificates at a B2 level, but is there any guarantee that they're actually at that level?

Linking a test score with an external measure such as the CEFR requires a standard that is somehow fixed or uniformly interpretable. Even if all tests link to the CEFR in the most thorough ways, its levels could only serve as a common currency when the same levels roughly mean the same thing to different users. Only two respondents unequivocally assumed that the B2 level is operationalized uniformly throughout Europe, but others were not so sure.

R6 It's very hard to say. I really do hope so. I hope that all tests are testing different strategies. I hope that we all are testing oral communication in a good way. I hope so, I hope so!

DISCUSSION

There is no uniform policy within Europe when it comes to language requirements for foreign L2 students who wish to study in a national language. In seven of the surveyed contexts there is a national regulation that specifies the official university entrance language requirements for L2 students. In the other cases there is either no language requirement whatsoever, or each university sets its own language level requirements. This diversification is reflected in the testing policy. In twenty contexts multiple language tests are accredited for granting university access, and in most of these cases centralized tests are accepted alongside local tests that have been developed by the accepting institutions themselves. Quite a few respondents doubted the quality and comparability of these local tests, and if there is one common wish the respondents share for the L2 university entrance policy in their context, it is increased standardization.

Throughout Europe universities have a lot of autonomy in setting the entrance requirements. Usually they are free to determine the language proficiency level they require for admission, although in some contexts the government may set some general requirements. In seven contexts, universities have no autonomy in deciding on the required language level and are obliged to follow governmental decrees. Irrespective of which body is responsible for determining the entrance level however, the reasons for choosing a certain language requirement appear to be quite unfounded. In only one of the 23 contexts where university entrance is determined by CEFR-linked language requirements, the required level is based on an empirical study. In 22 others, those requirements are not or only partly based on an analysis of the target language use context. Most often the language level requirements are determined by what other countries or universities do, or by the text of the CEFR itself. Moreover, in about one third of the contexts surveyed it is not uncommon for institutions to lower the linguistic entrance requirements to attract more students, or to manage the access of students with less desirable profiles. Since many institutions can decide on the entrance level they require, they are both policy makers and stakeholders at the same time, which leads to situations in which economic considerations may overrule actual student needs.

This study shows that the CEFR has fundamentally impacted university entrance language testing in Europe, and the most influential aspects of the CEFR are the six levels and the illustrative scales. In just one of the surveyed contexts the university entrance language requirements are not CEFR-related. This study confirms that the level most commonly used for university admission in Europe is B2, even though the respondents were not convinced that B2 is operationalized in the same way throughout Europe. Half of the respondents did not feel comfortable considering B2 as the default starting requirement for university. All

in all, however, the respondents of this study are rather positive about the CEFR, since they feel it offers a common – albeit sometimes vague – standard that has improved test score comparability.

The results of this study show that sometimes CEFR levels are interpreted very rigidly in a process mirroring the reification Fulcher (2004) warns against: B2 is used because it is B2, not because it is the level that best suits the users' needs. This study offers little support for the claim that the CEFR stimulates discussion between decision makers and language experts (North, 2014a). No respondent mentioned the CEFR as a catalyst for conversations between policy makers and test developers, and only one respondent claims that the university entrance level was based on a needs analysis. In the short term this kind of CEFR misuse can be considered unfair – since it does not offer every student equal opportunities across contexts – or irresponsible, since it ignores user needs or target language use demands in favor of norm-driven labeling. In the long run it could prove potentially destructive for the CEFR, since the levels might lose credibility. The CEFR has always been an open source hermeneutic, but in many contexts it now serves as a self-administered seal of quality. It can give university admission officers a semi-objective tool to control university entrance and it may allow test developers to claim a link to a certain level without having to offer any kind of proof for this.

CONCLUSION

The data presented in Chapter 1 indicate that linguistic university entrance requirements in Europe do not appear to be based on robust empirical foundations. Similarly, the B2 requirement, in spite of its prevalence, commonly lacks robust backing. This chapter has also shown that only one third of the respondents considered the B2 level as sufficient for international L2 students to function linguistically at the start of university. Other respondents judged that B2 was the absolute minimum, recommended offering B2 students additional language support, or considered B2 users insufficiently proficient to meet the linguistic demands of academic studies at university. Collecting empirical evidence regarding the performance of international L2 students in the target context is one of the goals of the following chapter.

CHAPTER 2
CONTENT & LEVEL REPRESENTATIVENESS

The previous chapter investigated the language requirements for international L2 students in university entrance policies across Europe, and reported a widespread, but often unsubstantiated use of the B2 level. In this chapter we focus on the Flemish context, and assess evidence for the B2 requirement based on data collected from international L2 students ($L2_P$ and $L2_F$) and Flemish university staff. This chapter not only investigates the B2 requirement: substantial attention is also devoted to examining how representative the ITNA and STRT tasks are for language use in the target context of Flemish universities.

Kane's (2013) Interpretation/Use Argument (IUA) is central to this chapter. Kane's thesis is that validity is not a property of a test per se, but of the interpretations or claims made on the basis of a score. Validation implies empirically determining the extent to which real-world test score uses or claims provide solutions for specific problems within a specific context (Gorin, 2007; Kane, 2013). Kane demands strong unequivocal evidence for claims, and when the stakes are high, he does not allow for any data that contradict a claim. Kane assigns great importance to score use, but he does not absolve test developers of all responsibility. Unless tests are used for purposes that clash with the intended use, test developers are responsible for claims pertaining to task selection, rating, and the like (see Figure 1.2).

ACADEMIC LANGUAGE REQUIREMENTS

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): argumentation, logic and analysis are considered central to the LAP construct, as is the ability to combine different sources and skills (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). Some authors have identified giving presentations, describing graphs, understanding lectures, summarizing texts and building an argument as prototypical LAP tasks (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 on whether a higher level might be more appropriate remains undecided (Taylor & Geranpayeh, 2011; Hulstijn, 2011; Xi et al., 2013).

Helpful as the descriptions of LAP characteristics above may be, they are perhaps too generic to provide detailed specifications for a university entrance language test. A general description does not suffice as the basis for a test that will be used within a specific context for a specific purpose (Lado, 1961), and local conventions may override general principles (Fløttum et al., 2006). Moreover, since most analyses of academic language proficiency relate to the Anglo-American tradition (Xi et al., 2013), their results will not necessarily apply to other contexts. Furthermore, the LAP characteristics described above characterize the language proficiency of an accomplished user of the academic idiom, not necessarily the language skills required of students embarking on their university studies.

JUSTICE

Since its earliest traceable origins the purpose of centralized testing has been to select individuals who possess a certain set of skills that are deemed important in the light of 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), examinations have very little to do with impartiality. In the Foucauldian tradition a test is considered an instrument of power that allows an in-group to select members from an out-group. Foucault's views have inspired language testers to critically examine the impact of tests on people's lives, and to question the gatekeeping functions they often perform (see Shohamy, 2001 for a seminal contribution). Recognizing the power imbalance inherent to testing, language testing organizations have developed a set of principles to ensure that test developers do not engage in activities inimical to candidates' best interests (e.g., ILTA, 2000). Set against this background of test ethics, this chapter investigates to what extent STRT and ITNA can justifiably serve as gatekeepers to university entrance.

Most post-Messick validity theorists, including Kane, have highlighted the importance of considering the social consequences of a test. In the wake of this, there has been an increased attention for issues of fairness and justice among language testers (Davies, 2010; Kane, 2010; Kunnan, 2010; McNamara & Ryan, 2011; Xi, 2010). While there is general agreement that fairness primarily concerns bias and impartiality (McNamara & Ryan, 2011), justice has proven more elusive, though some consensus does exist. Contrary to fairness, which always presupposes the existence of a test, justice questions the legitimacy of having a

test as a gatekeeper in the first place (McNamara & Ryan, 2011): In some cases the very fact that there is a test may introduce imbalance or inequity in a larger population (Kunnan, 2000). This idea is particularly relevant to the current chapter, which relates to a context in which one subpopulation (i.e., international L2 students) is required to pass a test before gaining entrance to an institution that is open to others (i.e., students with a Flemish secondary school degree).

While it is not the purpose of this chapter to fully conceptualize justice from a philosophical perspective, it aims to contribute to the debate on justice in the language testing literature by applying insights from the major justice theories and to use them as complementary to the IUA. Kane himself does not explicitly mention justice, but it is a logical extension of his sixth main statement: "the evaluation of score uses requires an evaluation of the consequences of the proposed uses; negative consequences can render a score use unacceptable" (2013, p. 1).

Much of what has been written about justice in language testing has been influenced by the writings of 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. Conversely however, 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), while 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 by arguing that L2 university entrance tests should not include tasks that L1 speakers are not expected to perform in the target context. But, as can be deduced from Rawls' 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).

Rawls' theory relies on the presumption that true justice exists. This idealistic approach obstructs its applicability in the real world. Sen (2010) therefore proposes an alternative theory in which justice is not seen as absolute, but as context-dependent. Sen's theory of justice is grounded in Rawlsian principles, but relies on equality of freedom and on the absence of injustice. If a situation is perceived as unjust, and freedom is restricted without a reasonable, rational argument, that situation *is* unjust. Dworkin (2003, 2013), a Rawlsian proponent of distributive justice, also supports the importance of freedom in his theory of justice. To Dworkin, any institution, large or small, is under the moral obligation to ensure equality of opportunity for all its members – also when this

implies unequal treatment. Rawls does not offer many practical guidelines for investigating justice, but his and Dworkin's work offer principles against which the justice of a university entrance policy can be evaluated. 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, it could be argued, with Sen, that justice can be defined 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 do not stand to reason or are unsupported by empirical data.

The Flemish university entrance policy limits the freedom of access of L2 students based on the assumption that students below a certain language proficiency level cannot successfully participate in academic studies. If this policy is just, people who fail the language test would not perform well in the target language use (TLU) context. If they did, 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 any would dispute the injustice of a policy that works to the disadvantage of people in an already disadvantaged position, yet lacks rational or empirical grounds.

RESEARCH AIMS

This study examines Assumptions 1 and 2:

A1 *B2 is an adequate threshold level to decide on international L2 students' access to a Dutch-medium university in Flanders.*

Based on the findings of Chapter 1 and on a comparison of the B2 descriptor with the LAP literature review, it was hypothesized that B2 learners would struggle to meet those linguistic demands.

A2 *STRT and ITNA are representative for the academic language requirements at Flemish universities.*

The hypotheses relating to Assumption 2 were that (a) the oral components of STRT and ITNA would be representative of the target setting, because they contain typical LAP tasks and that (b) ITNA's computer component would be less representative because it lacks productive writing - a crucial LAP skill.

The third aim of this chapter is to assess the justice of the Flemish university entrance policy for international L2 students. This requires determining whether freedom of access is restricted on grounds that are supported by empirical data or rational argumentation (Rawls, 2001; Sen, 2010).

PARTICIPANTS & METHODOLOGY

The data presented in this chapter include the perceptions of 24 academic staff members and the experiences of 31 L2 students regarding the linguistic demands of university studies. It draws on insights from needs analysis (Long, 2005; Gilabert, 2005) and mixed-method research (Creswell, 2015), and revolves around the triangulation of sources and methods by using a concurrent design in which quantitative data add an interpretative layer to qualitative data.

Participants

L2 students

The study is based on two groups of L2 participants representing the main research traditions (humanities, exact sciences and social sciences) at the three largest universities in Flanders. Some attended class in 1000-seat auditoria, while others did so in smaller groups of about 50 students.

Group 1 ($L2_P$)
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 twenty (range: 18 – 45), the median length of L2 Dutch instruction was fourteen months (range: 9 – 48) and most participants (7) were female. Three of these participants had entered Ghent University at bachelor level, eight at master level. These respondents were recruited after they had registered, so they had already passed a language test (ITNA = 10, STRT = 1). These participants will be referred individually by their pseudonym, or collectively as $L2_P$ (see Appendix 3).

Group 2 ($L2_F$)
In the summer of 2014, 135 non-native speakers of Dutch who planned to enroll at a Flemish university sat both ITNA and STRT as part of a concurrent validity study. Of this group, 68 candidates passed ITNA or STRT, granting them access to university. Less than half of the group (32) went on to register for a Dutch-

medium program at university. Twenty of these registered students agreed to participate in this study. Before the start of the academic year, one student decided to postpone her studies for financial reasons, so twenty participants remained (see Appendix 4). It is important to stress that these twenty participants all took both tests and seven of them received a different pass/fail outcome on STRT and ITNA. Because these students (except Stella – see below) had passed one of the two tests, they were allowed to register for university despite having failed the other entrance test. Quite likely, this is the first study to bypass the truncated sample problem (Wall, 1994) in such a way. This classic sampling problem entails that students who do not pass an entrance test cannot enter university, as a result of which there is no way of knowing how they would have fared.

Ten participants were freshmen, ten were master students. Six attended Ghent University, six were at the University of Leuven, and four at the University of Antwerp. Leila attended an interuniversity program. Stella had failed STRT and ITNA, but had been 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 (n = 17) was female. These participants will be referred to collectively as $L2_F$, or individually using their pseudonym. Appendix 4 provides additional information.

Data collection was carried out between October 2014 and July 2015. In any longitudinal study, attrition occurs and by the end of the data collection, seven students had left the project. Oeéana and Clara had dropped out in February 2015 to pursue studies in their L1. Anastasia quit in the same month because she had lost all motivation to pursue her studies. Stella (April 2015) and Yazdan (November 2015) had to give up because of visa issues. Jessica and Chloé 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 students are expected to meet 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 Ghent University, 12 KU Leuven) and the main academic traditions (6 humanities, 7 exact sciences, and 7 social sciences), both at professorial (15; 7 of whom were also directors of educational affairs) and at tutor (6) level. Four participants worked at the central administration (2 language

policy and 2 educational affairs). At the time of data collection, the majority of the participants had gained substantial professional experience at university (experience at university: *Md*: 22 years, range: 3-35) and in teaching at university (participants not working in administration, teaching experience: *Md*: 19 years, range: 3-29; experience with first-year students: *Md*: 16 years, range: 3-29). These participants will be referred to as Ac1 – Ac24 (see Appendix 5).

Even though there were no direct professional ties between participants of the same focus group, hierarchic differences do exist, and power issues can make individuals change their views to match group consensus (Reybold et al., 2013). For that reason each focus group began by collecting the individual opinions of each participant in a paper-based questionnaire (Kahneman, 2011), which formed the basis of the group discussion.

Data collection & analysis

L2 Interviews & focus groups

All interviews and focus groups were conducted by the author, who used a series of recurring must-ask questions but was free to elaborate on salient subthemes that emerged during the talk. The data were audio recorded and transcribed in Dutch, but specific quotes were translated into English.

The L2 interviews were conducted to determine whether L2 students who passed a B2 test felt ready for the linguistic demands of university (Assumption 1) and whether the academic language tasks they received in real life matched the ones operationalized in STRT and ITNA (Assumption 2).

The interviews of $L2_P$ 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. $L2_F$ participants were interviewed during the academic year 2014-2015. Their perceptions of the linguistic demands of university and of their own language proficiency in relation to those demands were a vital part of each interview, which focused on a different topic every time: the first weeks of university (October), classroom experiences (November), the first exams (February) and the students' social network (March). The April interview was replaced by a retest of STRT and the interviews in July looked back on the past year.

The purpose of the focus groups with the academic participants was to come to a cross-disciplinary consensus (Belzile & Öberg, 2012) concerning the linguistic demands that students are expected to meet at the start of university, and to assess whether these demands matched the B2 target level of the tests (Assumption 1). When a focus group began, participants were asked if they wished to differentiate between linguistic demands for L1 and L2 students, but all

agreed that the minimal linguistic demands had to be the same for all students, irrespective of their L1. Participants were then asked to individually estimate 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 3.1), which they rank ordered in terms of difficulty or ability. It is worth noting that 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 et al., 2011). Table 3.1 presents an overview of the samples used in the focus groups (excluding the speaking samples, since they will not be referred to specifically in this article) and identifies the source, the topic, the length, the percentage of low-frequency (≥ 5000) and high-frequency words (≤ 2000), the difficulty (as measured by Flesch-Douma, FD), and the 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 of a first-year (W2) and final-year (W4) student. The authentic reading sample (R3) was taken from the first chapter of a first-year sociology course book and was considered representative by the focus group members. 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. Given the demands of academic listening (Field, 2011), it was unlikely that the threshold level would be placed at B1. Consequently, to avoid ceiling effects, two C2 samples were included. Samples W2, W4, R3, Li1 and Li3 were linked to the CEFR by four experienced members of the above-mentioned committee.

The transcriptions were coded a priori and inductively (Dey, 1993; Miles & Huberman, 1994) using NVivo 11 For Mac. 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 earlier 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 $L2_F$ interviews conducted in

November 2014, using the a-priori coding scheme (for inter-coder agreement, see Table 1.4).

Table 3.1. 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	STRT performance	Advertising	186	83,2%	7,3%	52	
C1	W2	1st 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 (STRT)	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 (STRT)	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

Note. 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

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 they believed students should possess upon university entrance. Since the view of academics on these matters may differ from the perception of students, the $L2_F$ informants received the same questionnaire in February 2015. The views of the academic participants and the opinions of the $L2_F$ informants were used to assess how representative the task selection in STRT and ITNA is for the actual linguistic demands at Flemish universities (Assumption 2).

The list of language skills used in the questionnaire was based on the literature review above and on a list of commonly occurring task types in thirteen 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 that featured in at least seven of these thirteen tests were added to the list. Participants were free to add skills to the list, which happened on one occasion ("accurate expression of ideas"). The categories in the list (Table 3.3 shows the final version) 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 $L2_F$ student's choosing. Eleven lecturers gave permission to have their class audio recorded. Before, during and after the classes the researcher also took field notes.

These data were used to compare test tasks to real-life language tasks (Assumption 2). Additionally, the first 30 minutes of each class recording were transcribed and analyzed for word frequency (using *TST Centrale*, a lemma-based corpus for Dutch) and speed (words/minute) and compared to STRT audio prompts, allowing for a comparison between the linguistic demands of university lectures and the language demands of STRT audio prompts (Assumption 1). Lastly, the field notes were analyzed for instances that showed whether or not a participant was able to cope linguistically during class (Assumption 1).

Academic score transcripts & test/retest scores

In April 2015 the remaining $L2_F$ participants (N = 15) took two STRT test tasks again: writing a summary of a scripted lecture about industrialization and giving a ten-minute presentation about pollution, based on slides. Since practical reasons prevented administering 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 β = .52, p < .000 presentation β = .57, p < .000).

At the end of the second semester (July 2015), the $L2_F$ 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 the courses they had taken up ($L2_F^+$, N = 8), and those who had not ($L2_F^-$, N = 8). The $L2_F^-$ group did not include the two students who had left university because of immigration problems, or the two students who had left the project early. The academic performance data were combined with the entrance test results in order to assess whether any academically successful L2 students had failed STRT or ITNA, which implies that they would not have been able to register for university if they had taken only that particular test (justice).

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 $L2_F^+$ students had achieved higher initial STRT or ITNA scores, and to measure score gains on STRT tasks. Since the tests' CEFR-based scales may be too broad to measure gains over a period of eight months, more detailed analyses were conducted, based on a methodological approach adopted by Serrano et al. (2012) and Llanes et al. (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. The results will be referred to below, but the analyses themselves are discussed in detail in Chapter 6. All quantitative analyses were conducted with *R* (*QuantPsyc* and *car* packages).

RESULTS

Assumption 1 **B2 is an adequate threshold level to decide on international L2 students' access to a Dutch-medium university in Flanders.**

The data used to examine this assumption are:

- test/retest scores, to measure differences in L2 proficiency over time;
- Focus group discussion data about the listening, reading and writing samples (see Table 3.1), to determine the minimal level of competence the academic staff members expected. Speaking samples were not part of the discussions, since all groups agreed that it was the least important skill for first-year students to master;
- Interviews with L2 participants, to cross-check the focus group results and to provide concrete examples of the linguistic hurdles they faced;
- Field notes, to provide first-hand observations of how L2 participants experienced lectures;
- A comparison of the lexical demands and speed of actual lectures and STRT listening tasks, to determine whether the participants' perceptions were confirmed by actual observations.

Listening

The focus group participants rank ordered the samples (see Table 3.1) before determining the minimally expected level. For listening, reading and writing, the focus group participants' rank order of the samples aligned with their CEFR levels.

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 group 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 $L2_P$ participants struggled to understand the natural, unpolished language of university lectures.

> 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.
>
> (Noor, October 2012)

Some L2 participants dropped out (Noor), quit going to classes (Océane, Clara) or experienced loss of motivation (Hoang, Merveille) primarily because they had problems understanding lectures. Most $L2_F$ participants felt unprepared for the listening demands of university lectures, and 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).

> [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?*
> 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?*
> Seriously. Nothing.
>
> (Océane, October 2014)

At the end of the year, seven $L2_F$ participants felt quite sure that they understood classes better than at the start of the year, although unfamiliar accents or unclear pronunciation remained a persistent problem for most.

> 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.
>
> (Merveille, June 2014)

The interviews showed that lexical problems caused additional difficulties during lectures. A comparison (see Table 3.2) between the language used in eight scripted lectures used as STRT prompts and in twelve actual university lectures confirmed this: authentic 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.

Table 3.2. University lectures and STRT listening prompts

		1K-2K*	5K-7K	7K+	w/m♯
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

Note. * % of words used in frequency band
♯mean words/minute

The field notes reveal other, more qualitative differences between the test prompts and in-class experiences. All bachelor and master 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: "there is a constant buzz of students talking among each other during class. The professor just talks through the noise" (Field notes Alexandra, 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 Alireza, 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 represents the actual language of syllabi. In the

end the consensus was that it is unrealistic to expect students entering university to cope with texts of this level, although it is the kind of language they will encounter early on in their studies. The groups finally decided to expect students to master R2 (C1) at the start of university, but not R3, because of its structural complexity.

In every focus group text R1 (B2) and text 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). The main reasons why both texts were considered too easy had to do with their clear structure, low information density and comparatively simple development of ideas.

For the $L2_F$ participants reading presented a problem, but one they said they mostly managed. All $L2_F$ participants reported that reading took much longer in Dutch than in their L1 because they looked up words, because they consulted sources in English or in their L1 to understand concepts they did not grasp in Dutch, or because they translated parts of their courses. At least three participants had translated their entire courses into their L1.

> In all honesty, I'm a bit of a maximalist. […] I lose a lot of time by translating.
>
> *I* *Do you translate your courses?*
>
> 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.
>
> (Stella, February 2015)

As the year progressed, quite a few $L2_F$ participants reported a perceived improvement in terms of reading comprehension (Gabriela, Emma, Océane, Clara) or speed (Alexandra, Marie, Stella). Other participants (Guadalupe, Hoang) confirmed that their reading had improved, but was not up to standard yet.

> There is one book about stuff Freud wrote – very difficult language […] I try to read it, but do not understand it.
>
> (Guadalupe, June 2015)

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 $L2_F$ 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. Other students had not yet received a writing assignment and had only taken multiple-choice exams. Quite a few $L2_F$ participants did not assume that their written skills had improved since the start of classes. Some even felt that their written Dutch had gotten worse (Anastasia, February 2014).

Speaking

There was overall consensus that for first-year students, receptive skills are more essential than productive skills, and that speaking is of little importance: "Speaking just does not happen in the first year [...] First and foremost, students entering university should be able to store information" (Ac 4). The university staff respondents were asked to determine a cut off score for the skills they considered important. Since speaking was considered the least important skill for students to master when they start at university, no minimum proficiency level was determined.

> Ac7 We can keep our expectations low, cause they don't need to speak in the beginning, do they? [The B2 sample] is definitely going to survive at university.
> Ac8 Basically, no faculty has any real demands when it comes to speaking.
> Ac6 True. Maybe we can lower the expected level then.
> Ac8 The cut off point could go below [B2] then.
> Ac6 Or maybe just above?
> Ac7 Look, do we need to test it at all?
> Ac6 Yes, you're right.

After two months at university, four $L2_F$ participants reported speaking Dutch quite often. Others had rarely used it (5), were afraid to use it (5), or had not

spoken Dutch yet (4). Likewise, 10 of the 11 $L2_P$ 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 students found ways to avoid speaking here too, by using chat (Janet) or e-mail (Leila) to contribute to group discussions.

> 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.
>
> (Leila, November 2014)

Leila 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 $L2_F$ interviews, but they are beyond the scope of this chapter. These themes will be discussed in Chapter 6.

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: $W = 31$, $p = .159$, $r = -.31$; 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 (see Chapter 6) 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). Since 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 $W = 48.5$, $p = .505$, $r = -.156$).

Assumption 2 **STRT and ITNA are representative for the academic language requirements at Flemish universities.**

The data used to examine this assumption are

- The L2F participants' opinion of the tests' representativeness;
- The experiences of L2P and L2F participants;
- The results of the academic language skill questionnaire in the focus groups and in the L2F interviews.

In October 2014, when asked which test they preferred, six $L2_F$ participants chose ITNA, ten 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 in ITNA. ITNA's least useful tasks according to seven participants were the vocabulary tasks, because they were perceived as unrepresentative for the vocabulary used at university.

> In practice we never use proverbs, but we do sometimes hear them. Many of the words I studied for ITNA I have forgotten. When you don't encounter a certain word at all, you forget it.
>
> (Marie, June 2015)

The $L2_F$ 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.

The interviews with $L2_F$ participants and the university staff focus groups clearly showed that all respondents judged receptive skills to be the most important. For students entering university, productive skills are less important, and speaking is a rare requirement:

> 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.
>
> (Heddi, December 2012)

Having established the relative importance of receptive and productive skills, the university staff participants took the questionnaire to decide which academic language skills are most important for first bachelor students to master when they enroll at university. Table 3.3 below indicates how essential each focus group considered each academic language skill.

Table 3.3. Academic language skills selected in focus groups (*N* = 6)

	#	+	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		
Understand general academic lexis	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					

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

The consensus in every focus group was that when it comes to speaking or writing, *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 of primary importance, their selection of essential academic language skills also included skills that are important for passing written examinations. Writing down answers in a meaningful, accurate and structured way matters a great deal in that respect.

When the $L2_F$ participants received the questionnaire in February 2014, their selection reaffirmed the importance of receptive skills: "*understand general academic lexis*", "*understand implicit message*" and "*understand scientific text as a whole*" were most often identified as important. "*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 it is the lecturer's task to introduce this. Every $L2_F$ informant on the other hand complained of limited lexical knowledge as a major obstacle to attending classes. In most cases, $L2_F$ participants were not actually referring to highly specialized terms, but to words that are commonly acquired in the course of Flemish secondary education. Possibly, the university staff underestimated the lexical complexity of their own language use, 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, Alexandra was unfamiliar with basic mathematical terminology at the start of university.

> 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.
>
> (Alexandra, October 2014)

"*Understand implicit messages*", was also perceived differently by academic staff and $L2_F$ students. Professors were convinced that "academic language is not supposed to be implicit" (Ac 7), but for $L2_F$ students, implicit language includes irony, jokes and idioms – all of which are important when attending lectures. During these lectures, most $L2_F$ 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 $L2_F$ participants wrote "comments on a hand-out" (Ac22) without taking actual notes.

At least as important as the skills the participants selected, are the ones they did not select. All $L2_F$ participants and all university staff members disregarded "*give a presentation*" and "*describe graphs and tables*". Nevertheless, delivering an oral presentation is one of the two tasks included in the oral components of STRT and ITNA, and at least two STRT tasks rely on candidates being able to describe graphic or tabular input. All this shows that the test content differs from reality in a number of aspects; the next section of this chapter focuses not so much on content, but on level requirements.

The justice of using ITNA and STRT as gatekeepers to university admission

Because of the specific design of this study, in which all $L2_F$ participants took both tests, candidates who failed one test but passed the other could still register for university. The following sources of data were used to assess matters of justice:

- The participants' perceptions about the justice of the university entrance policy;
- The initial STRT and ITNA outcomes, presented in appendix 3;
- Indicators of academic success (i.e., $L2_F^+$ and $L2_F^-$ 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 since there are virtually no support systems for L2 students (Ac24), and since they "are in the auditoria with other students [and] it's better to give these students a clear message from the start" (Ac17). Most $L2_F$ 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 $L2_F$ 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" (Clara).

The academic results of the $L2_F$ participants seem to partially confirm Clara's point: there is no clear link between language test scores and academic success. $L2_F^+$ students did not significantly outperform $L2_F^-$ students on the initial STRT and ITNA tests (STRT: $W = 46$, $p = .625$, $r = -.115$; ITNA: $W = 51$, $p = .599$, $r = -.120$). On the STRT retest too, $L2_F^+$ did not outperform $L2_F^-$ (STRT writing: $W = 14$, $p = .741$, $r = -.104$; STRT speaking: $W = 16$, $p = .451$, $r = -.238$). When interpreting these outcomes, is important to note that only thirteen $L2_F$ participants took part in the final exams of that year.

Participants who were academically unsuccessful yet gained admission on the basis of a language test can be considered false positives, in the sense that they gained entrance to university but were, for various reasons, not able to successfully complete the first year, whereas participants who failed STRT or ITNA yet belonged to the $L2_2^+$ group, are false negatives. Of the sixteen participants who did not drop out or involuntarily leave the project, STRT and ITNA respectively assigned seven and six false positives. Since false positives do not lead to exclusion of members of a specific group however, they do not qualify as an injustice. From a justice perspective, false negatives carry considerably more weight. In the $L2_F^+$ group, ITNA assigned two (Leila, Guadalupe) false

negatives, STRT none. Leila was not a confident speaker, but passed her exams with honors. Guadalupe had experienced a difficult first semester, but passed all of the second semester exams.

DISCUSSION

The university staff participants agreed that L2 students would inevitably be less proficient than their L1 peers 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 L2 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 reached a language proficiency level that matches the linguistic demands of the TLU context. The results show that – especially for listening – this is not the case.

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 $L2_F$ 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 the listening demands of university. 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; Ranta & Meckelborg, 2013).

The fact that most participants reported listening as the most problematic skill does not imply that they were proficient enough 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 would in their L1, or had to translate coursework to their L1. $L2_F$ 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 ties in with the justice discussion below.

This study offers little – if any – data to support the assumption that students who pass the B2 language test are able to cope with real-life linguistic demands of academic studies. All L2 students included in this study had passed ITNA or STRT or both (except for Stella – see above). Some managed remarkably well, but the majority of L2 participants was not ready to deal with the linguistic demands of academic studies at university (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.

The second research goal of this study was to investigate Assumption 2. Here, the results show that to some extent STRT and ITNA appear representative of the communicative demands of academic programs at Flemish universities. STRT takes into account content-related criteria, which corresponds to the importance the university staff assigned to meaningful rather than correct language. ITNA only considers linguistic correctness, however. 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 $L2_F$ participants. The $L2_F$ participants and the university staff agreed on the importance of argumentation and note-taking. Both are operationalized in STRT, but the operationalization of note-taking does not truly take into account trends in Power Point-based teaching (Lynch, 2011).

In some cases, however, the operationalization of STRT and ITNA contrasts with real-life demands. The university staff participants and the L2 students at bachelor and at master level agree that for students at Flemish universities receptive skills are more important than productive skills. Oral skills are considered to be of the least importance. Strikingly, all $L2_F$ participants and all university staff members did not consider giving a presentation or describing graphs and tables to be important skills. This does not necessarily mean that productive skills should not be assessed in university entrance tests, because oral skills will likely impact 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 bachelor 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 well 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 necessarily taking into account the specific features of the Flemish context for a representative selection of test tasks.

Apart from examining evidence regarding Assumptions 1 and 2, this study questioned the justice of using ITNA and STRT as gatekeepers to university admission. Carlsen (2017) 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 manage well at university. Since 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. 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.

In essence, 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 often is impossible to trace false negatives. In the design of this study however, the problem of truncated samples (Wall, 1994) was bypassed by tracking seven $L2_F$ participants who had actually failed STRT or ITNA. ITNA assigned two false negatives, STRT none. If Stella – who had a good academic record but left the country due to visa problems – is included in the count STRT and ITNA respectively assigned three and one false negatives. False negatives signal an unfounded restriction of access that applies to one subpopulation alone (Kane, 2013). According to leading justice theorists, this might shed doubt on the justice of an entrance policy (Rawls, 1971, 2001; Dworkin, 2003, 2011; Sen, 2010).

CONCLUSION: ASSUMPTIONS 1 AND 2

The results of this study reveal that L2 students who passed ITNA, or STRT, or both, were not ready for the receptive linguistic demands of academic studies at university (Assumption 1). It is also safe to conclude that the content of the Flemish university entrance tests at points deviate from real-life language demands (Assumption 2). The observation that a number of students who were assessed below B2 actually managed at university, also qualifies as negative evidence regarding Assumption 1. Requiring an even B2 level does not appear to be a very effective way to discriminate between students who are likely to manage the linguistic TLU demands, and those who are likely to struggle.

PART 2
SELECTION & DISCRIMINATION

The first part of this dissertation investigated two aspects of the Flemish university entrance policy: the B2 level requirement (Assumption 1), and the representativeness of STRT and ITNA for the target context (Assumption 2). Chapters 3 – 5 included in this second part focus on test equivalence (Assumption 3) and the language proficiency level of first-year students with a Flemish secondary school degree (Assumption 4).

Chapter 3 examines whether STRT and ITNA can be considered equally difficult, and whether comparable tasks measure similar constructs. Next, Chapter 4 scrutinizes the part of STRT and ITNA for which direct one-on-one comparisons can be made most directly: the linguistic criteria used to score oral performances. The outcomes of both chapters indicate that STRT and ITNA cannot be considered equivalent in terms of difficulty, construct, or rating scales.

Assumption 4 – on the language proficiency level of Flemish high school graduates – is one of the research questions of Chapter 5, which also investigates performance differences between the performances of two groups of L2 learners. The results quite clearly show that not all students who graduated from a Flemish secondary school are likely to attain the B2 level.

Chapters 3, 4, and 5 are based on:

Deygers, B. (2017, in press). University entrance language tests: examining assumed equivalence. In J. Davis, J. Norris, M. Malone, T. McKay, & Y Son (Eds.). *Useful Assessment And Evaluation In Language Education. Washington*, D.C.: Georgetown University Press.

Deygers, B., Van Gorp, K., & Demeester, T. (2017, in press). The B2 level and the dream of a common standard. *Language Assessment Quarterly*.

Deygers, B., Van den Branden, K., & Peters, E. (2017). Checking assumed proficiency: Comparing L1 and L2 performance on a university entrance test. *Assessing Writing*, 32, 43–56.

These papers have been edited to fit the structure and approach of this book. Sections pertaining to the research context, the methodology, and the participants have been revised to avoid redundancy.

CHAPTER 3
LEVEL & CONSTRUCT EQUIVALENCE

> The first chapter showed that universities often require prospective international students to pass a language test as a precondition for admission. In Europe, the most commonly required language level for this purpose is B2. Quite often, different tests are accepted as proof of B2 proficiency, but usually without empirically determining the relationship between the different tests.

This chapter examines whether STRT and ITNA can be considered equivalent measures of Dutch language proficiency at B2 level. If this assumption (A3) is true, or largely true, accepting either STRT or ITNA poses no immediate problem. But, if it is false, and if one test is substantially more difficult than the other, it could lead to an unjust entrance policy. In investigating whether the target level and the constructs of STRT and ITNA are comparable, this study draws on Kane's (2013) Interpretation/Use Argument and on Phillips's (2007) ideas on policy effectiveness. The implications of the findings are discussed with regards to justice (Kunnan, 2000; McNamara and Ryan, 2011; Rawls, 2001; Sen, 2010).

The implication of Kane's logic is that when scores of two different tests carry equal weight in a university entrance policy, university admission officers rely on an assumption that requires validation, since it has important social consequences for the candidates. In Kane's logic (Kane, 2013, p. 62, but see also Bachman and Palmer, 2010), since neither STRT nor ITNA have made any claims regarding their equivalence, this assumption is for score users to prove. To date, however, no empirical evidence has been offered in this regard.

EQUIVALENCE

When a university claims that the certificates of two different tests are equivalent measures of a certain level of language proficiency, and this claim is wrong or unsubstantiated, it may have serious consequences for the educational standards of a university and on the lives of test takers. First, when a university entrance language policy wrongfully assumes that different language tests measure an equivalent level of language proficiency, this policy may cause unacceptable variation in the assessment of the language abilities of the admitted student population, thereby failing to meet its main goal. Second, when two tests are assumed to be equivalent, test takers should have a comparable chance of passing either test. When test takers need to make a choice between two tests on the

basis of unreliable information, the justice of the university entrance policy can be questioned.

A number of studies have examined the relationship between scores on two university entrance tests by calculating correlations. Fulcher (1997), comparing a local test with TOEFL results, computed an overall correlation of r = .64. Another study (ETS, 2010) reports a .73 correlation between TOEFL iBT and IELTS, and correlations between .44 (writing) and .68 (speaking) for the subskills. The ETS researchers further investigated the relationship between the tests using regression-based prediction, equipercentile linking, and conditional probability. The report excluded the regression and conditional probability analyses because they were less informative than the equipercentile findings. Zheng and De Jong (2011) compared the PTE Academic test to English language tests that are used for university admission purposes, such as TOEIC (r = .76) and TOEFL iBT (r = .75), and used regression analysis to map PTE Academic scores onto the CEFR (r^2 = .5). Lastly, Riazi (2013), building on the aforementioned study, found an overall correlation of r = .82 between IELTS and PTE Academic scores, and correlations between r = .66 (listening) and r = .72 (speaking) for the four skills. Using t-tests, Riazi further showed that PTE Academic significantly differentiated between candidates who received higher and lower IELTS band scores. Riazi reported medium effect sizes for the productive skills (Speaking η^2 = .50; Writing η^2 = .50), and close to medium effect sizes for the receptive skills and the overall score (Listening η^2 = .38; Reading η^2 = .44; Overall η^2 = .45).

In sum, recent studies that examined the relationship between university entrance language test scores often show medium to strong correlations. All of the studies consulted provided correlation coefficients, and some (e.g., ETS, 2010) offered evidence from other types of analyses, because relying on correlational evidence alone can be rather misleading (Kane, 2013; Lissitz & Samuelsen, 2007; Norris, 2016). Usually, these studies were based on official score transcripts (e.g., ETS 2010), sometimes combined with self-reported data (e.g., Zheng & De Jong, 2011), but little or no studies have been published in which researchers have had access to detailed rater data from two different, high-stakes entrance tests.

Another important aspect of test equivalence is the extent to which two tests used for the same purpose in the same context measure similar constructs. If two tests that are assumed to be equivalent do not measure the same level of language proficiency, examining construct equivalence may shed light on the nature of this mismatch (Lindridge, 2015). Clearly, as was the case for examining level equivalence, using correlational evidence to tackle research questions relating to construct equivalence is insufficient (Kane, 2013; Norris, 2016; O'Loughlin, 2001; Shohamy, 1994). Correlations may be spurious (Lissitz & Samuelsen, 2007), they may hide underlying discrepancies (Harsch & Martin, 2012), and they do not offer information concerning the nature of the relationship between two tests. For that reason, in the social sciences, construct equivalence

research is often conducted using inferential statistics such as exploratory factor analysis (Welkenhuysen-Gybels & van de Vijver, 2001).

In language testing, however, little quantitative research has been conducted to examine construct equivalence, even though it can contribute to a deeper understanding of the outcomes of level equivalence research (Wang, Wang, & Hoadley, 2007). For score users, such as admission boards, students, or teachers, it could be informative to know why test results differ and where tests that serve the same purpose are actually dissimilar. Universities offering post-entry language courses for international L2 students could use this information for instructional purposes, for example. Or, programs with strict writing requirements could be interested in learning how ratings on equivalent tests relate to each other.

JUSTICE

Phillips (2007), observing that policy measures are not always founded on empirical data, recommends critically examining policy claims by identifying the problem that the policy was intended to solve, and by evaluating the effectiveness of the proposed solution on the basis of evidence. The decision rule (Kane, 2013; Phillips, 2007) quite simply states that a policy measure cannot be maintained if it does not solve the problem it was meant to address. Translated to the context of this study, the decision rule implies that, if the use or interpretation of a test score within a university entrance policy is unsupported by empirical data, there may be reason to doubt its effectiveness.

Kane (2013), referring to Phillips (2007), requires the evidence used to validate a policy claim be proportionate to the social consequences of that claim. And, when the stakes are high, all the evidence should support the claims made by score users (Kane, 2013). If empirical evidence does not support the way in which policy uses test scores (in this case, as equivalent), the policy may be ineffective, but it might also be unjust.

Justice, in this case, is concerned with the effect of introducing one or more language tests on a larger population. In line with Rawlsian (Rawls, 1971, 2001) logic, the prerequisite for justice is fairness, that is, freedom from bias (Rawls, 2001; Sen, 2010). But, even if all tests accepted in a university admission policy are equally fair, the admission policy can be unjust when it causes an indefensible disequilibrium in a population (Kunnan, 2000; McNamara & Ryan, 2011). When a candidate is more likely to get into university simply because of picking test A rather than test B, without being aware of a possible difference in pass probability, the university entrance policy may be unjust, since it may restrict freedom of access to university on grounds that are unsupported by empirical data (see Chapter 2).

RESEARCH QUESTIONS

The two research questions that guide this chapter tackle the assumption of equivalence (i.e., Assumption 3) from two angles:

RQ1 *Is there empirical evidence to support the claim that STRT and ITNA certificates are equivalent measures of Dutch language proficiency in the Flemish university entrance policy?*

Secondly, this chapter aims to explain the reasons for a possible mismatch by considering construct equivalence.

RQ2 *What is the nature of the relationship between comparable ITNA and STRT tasks, constructs, and criteria?*

The study described here contributes to the existing literature by explaining the results of level-equivalence research by means of construct-equivalence methodologies. In addition, the results are used to explore the ethical consequences of the university entrance policy with reference to principles of justice.

PARTICIPANTS & METHODOLOGY

The introduction of this dissertation offered further information on STRT and ITNA. Appendix 1 and 2 provide a detailed overview of the STRT and ITNA tasks. Nevertheless, it may be useful to reiterate that ITNA and STRT differ most in the written components, but contain highly similar oral tasks. The computer section of ITNA contains selected-response or gap-filling tasks, whereas STRT's writing tasks are integrated and require a lot of writing. Even though the operationalizations differ, the written components of ITNA and STRT are both scored for reading, listening, vocabulary, grammar, and cohesion. The oral components of both tests include a presentation and an argumentation task.

Performances on ITNA are scored after the test by two trained examiners who reach a consensus score for five linguistic criteria based on the candidate's performance on both tasks. STRT is centrally scored by two trained raters who score content criteria in a binary way (i.e., whether the candidate mentions the required aspects or not) and linguistic criteria on a four-point scale. The linguistic criteria in both tests are based on the same CEFR descriptors (*Vocabulary*, *Grammar*, *Cohesion*, *Fluency*, and *Pronunciation*), but the STRT

rating scale includes two additional criteria (*Register* and *Initiative*). Scoring on the ITNA computer test is binary and automated. As in the oral component, the written part of STRT is scored for content criteria and linguistic criteria by two independent, trained raters.

STRT candidates who achieve an overall Rasch measure at or above 1.42 (private communication, 6 January 2016) receive certification. ITNA candidates who score 54% or more on the computer test, may take part in the oral component. ITNA candidates who take the oral component and attain an overall score of 52.5% or more, get the B2 certificate.

Below, rating criteria will be printed in italics with a capital letter. To avoid confusion, *Vocabulary*, as a criterion will be printed differently than vocabulary, as a linguistic competence (using the CEFR's terminology).

Data collection

In order to determine whether the same candidates received comparable scores on STRT and ITNA, and – if not – why, test performance data of the same candidates on both tests were collected for the purpose of this study. From May 2014 through September 2014, all ITNA candidates (N = 802) were invited to take STRT free of charge. Since ITNA scores are communicated to the candidate within two days of taking the test (while it may take up to a month before STRT results are known), it was assumed that successful ITNA candidates would be disinclined to still take STRT. Thus, to avoid attrition, all participants first took STRT no more than one week prior to taking ITNA.

All written STRT tests were administered under the conditions prescribed in the examination manual, and in the presence of the researcher. Trained examiners conducted the oral examinations, following a procedure that is highly comparable in STRT and ITNA. The candidate receives the speaking tasks, gets preparation time (typically ten minutes) and returns to perform the task in front of the examiner. All participants also took the ITNA computer test under prescribed examination conditions. Trained ITNA examiners administered the oral ITNA tasks. All performances were scored by trained STRT and ITNA raters.

$L2_F$ participants

Between June 2014 and September 2014, ITNA candidates were invited to take the STRT free of charge one week before the ITNA administration, which granted them an extra opportunity to gain access to university. The predetermined stopping criterion for data collection was the start of the 2014-2015 academic year. After omitting incomplete performances (some participants gave up during one or both tests), 118 participants remained in the dataset that was used to compare the results of $STRT_{written}$ and $ITNA_{computer}$. Since ITNA candidates who score

below 54% on the computer test cannot take part in the oral component, the number of participants that could be meaningfully compared for the oral tests was reduced to 82.

The exams were administered at the largest Flemish universities (37% at the University of Antwerp, 34% at Ghent University, and 29% at the University of Leuven) by trained examiners. The first author was always on site to ensure the consistency of the test administration. Rating the STRT test takes a few weeks because all performances are scored centrally. However, the ITNA scores in the current administration were available on the day of testing. The candidates received no further formalized instruction between the tests, and given the one-week time span between the two administrations, it was assumed that their language skills remained constant. The performances were rated anonymously under normal rating conditions.

Table 4.1. Research population variables vs. regular STRT and ITNA populations

		$L2_F$		ITNA		STRT	
Age	Mean (SD)	27 (7)		28 (8)		26 (7)	
	Min - Max	16 – 50		15 – 61		14 – 60	
Gender	Female	70%		67%		65%	
	Male	30%		33%		35%	
Educational goal		66%		76%		58%	
L1		French	17%	French	16%	French	29%
		Spanish	8%	Spanish	11%	German	25%
		Arabic	7%	Arabic	9%	Papiamento	7%
		Russian	7%	Russian	8%	Dutch	6%
		German	6%	German	5%	Russian	4%
N		138		485		521	

The $L2_F$ participant population was representative for the population of both tests in terms of age ($L2_F$ $\bar{X}$ = 27, SD = 7; ITNA $\bar{X}$ = 28, SD = 8; STRT $\bar{X}$ = 26, SD = 7), gender ($L2_F$ 70% female; ITNA 67% female; STRT 65% female), and nationality. In terms of L1, the actual STRT population had a slightly different distribution due to the large number of candidates from Belgium's neighboring countries, expat communities, and countries with historical ties to the Dutch language. Table 4.1 displays key demographic variables for the full $L2_F$ population, and the typical STRT and ITNA populations.

T-tests showed that the distribution of scores found in the respondent population corresponds to the score distribution in the overall test populations. No significant differences were found between the final scores of the $L2_F$ population and those of the total ITNA population who took the test in the same period (t (485) = -.493, p = .622). For STRT, this study was the first administration of a new test version, and apart from pilot data, no other scores were available.

Levene's test for equality of variance (Field, Miles, & Field, 2012) confirmed the variance comparability of the final scores of the sample population to the regular STRT population in the last administration of the previous STRT test (F = 0.014, p = .907).

Data analysis

Level equivalence

In order to facilitate the interpretation of the descriptive statistics, the total raw scores of STRT and ITNA (172 and 125 respectively) were recalculated into a percentile scale. The significance of the difference between overall, written, and oral test results was determined using t-tests (Riazi, 2013). Cohen's d was used to calculate the effect size for the difference between test results for the full population (N = 118). Since it was assumed that both tests were likely to agree on the best and the worst performances, the scores within the interquartile range (i.e., the range excluding the 25% highest and 25% lowest performances) were examined as well. Wilcoxon's rank-sum test, a non-parametric test, was used to determine the significance of the differences within the interquartile range, and Cohen's d was used to estimate their magnitude.

Both parametric (r) and non-parametric (τ) correlations were used to describe the strength of the relationship between the overall scores, the written scores, and the oral scores. Interquartile correlations were used to measure the strength of the agreement around the cut-off point.

For university admission in Flanders, only one thing really matters, and that is getting the B2 certificate. In order to get a clear idea of the disagreement in terms of pass/fail judgments, a crosstab was constructed on the basis of binary overall STRT and ITNA outcomes. Furthermore, the probability of passing either test was computed. McNemar's binomial sign test served to determine whether the difference in pass/fail judgments between the two tests was significant.

Furthermore, in order to gauge the strength of the relationship between overall test scores, and scores on the written and the oral components, parametric and non-parametric correlation coefficients were computed for the full population and for the interquartile population. The Pearson correlation coefficient was used to assess the strength of the relationship between the total scores and between the scores on the written components. Because of the sample size (Howell, 1997), and because of considerations regarding restriction of range (Field, Miles, & Field, 2012), Kendall's Tau (τ) was chosen to correlate the oral and the interquartile data.

Construct equivalence

The ITNA and STRT speaking components contain the same task types and many corresponding rating criteria. Both scales are based on the same CEFR levels (A2 – C1), but for certain skills (e.g., *Grammar*), the ITNA scale differentiates between basic proficiency levels and plus levels, where STRT only has one level. When the ITNA rating scale included a plus level, but the STRT scale did not, both ITNA levels were merged (i.e., ITNA's B1 and B1+ both became B1, and ITNA's B2 and B2+ both became B2). All scores were recoded in collaboration with the coordinators of both tests to ensure that no interpretative errors were made. Any recoding was done after the performances were rated, so all raters used their own scales in the way they were trained to use them.

Since the rating criteria were known to be correlated, oblique promax rotation was used to run a Principal Component Analysis (PCA) on the standardized z-scores of the oral component. Since ITNA only offers scores on both tasks combined, the PCA was run using the STRT scores for both tasks combined. After having determined that the preconditions for a PCA were met (Bartlett's test of sphericity: $X^2(4) = 358$, $p < .000$, $KMO = .82$), the initial analysis showed that three factors had eigenvalues at or above 1, explaining 70% of the score variance. The scree plot was used for confirmatory purposes, and showed that a three-factor solution was warranted.

Linear regression was used on the written and the oral datasets to determine how well STRT ratings predicted ITNA scores (Zheng & De Jong, 2011). In both datasets, the number of cases with large residuals (written, 4%; oral, 4% after removal of two outliers) was within limits, Cook's distance was never >1, no individual cases were more than three times the average leverage, the covariance ratio was satisfactory, and the assumptions of independence and multicollinearity were not violated (Norris, 2015; Purpura, Brown, & Schoonen, 2015). A multiple linear regression analysis was conducted to determine how much score variance in $ITNA_{computer}$ was predicted by $STRT_{written}$ scores. For the oral test component, a regression model was constructed with the overall oral ITNA scores as a function of the STRT criteria scores. Given the limited sample size, we could not reliably assess the contribution of the predictors in the oral model, but we were able to generalize from the overall model fit (Field, Miles, & Field, 2012).

Regression analysis helps to see how different variables relate to one another, but they do not yield insights into the relative difficulty of tasks, tests, and criteria that Multi-Faceted Rasch analysis (MFRA) offers (Bond & Fox, 2007). MFRA considers a test score the result of an interaction between different facets, such as candidate ability and test, task, or criterion difficulty (McNamara, 1996). MFRA software, such as *FACETS* (Linacre, 2015), estimates the interaction between different facets that contribute to a test score and maps them onto a

common logit scale. Each facet is composed of different variables that are called elements. The difficulty of an element is called a measure, which is expressed in logits. Infit Mean Square (MnSq) statistics show how well an element fits the Rasch model and indicates to what extent the elements of a facet fit the same construct. The closer the Infit MnSq is to 1, the better the data fit the model; values below .5 indicate redundancy, and values above 1.5 are seen as misfitting or disruptive to the model (Barkaoui, 2014).

MFRA was used in this study to compare the relative difficulty of $ITNA_{computer}$ and $STRT_{written}$ (N = 118) and to rank the tasks on both tests in terms of relative difficulty. For this purpose, all tasks were weighted equally in the Rasch model. A second Rasch model was constructed using equally weighed oral criteria (n = 82) in order to determine how the criteria on both tests ranked in terms of difficulty. Lastly, a third MFRA used the real weights of the oral criteria to compare the actual difficulty of both oral tests relative to each other.

Prior to these analyses, instances of candidate misfit were examined. One case was removed from the dataset, as this candidate had prematurely ended the oral component of STRT, resulting in an incomplete set of observations (this candidate was removed from all analyses included in the current study). In the fifteen remaining cases of candidate misfit, it concerned disagreeing judgments on STRT and ITNA, which was relevant in light of the research questions. Since the dataset contained eighteen cases in which there was a 30% difference in raw scores (converted to the same percentile scale), it was decided to give preference to Infit MnSq measures rather than to Outfit MnSq measures, which are more sensitive to outliers.

RESULTS

Level equivalence

The overall STRT-ITNA correlation is strong (r = .767**), as is the relationship between $STRT_{written}$ and $ITNA_{computer}$ (r = .694**). Other correlations result in much lower coefficients, and considering the similarity of the oral components, the correlation coefficient of τ = .387** is striking. Low interquartile correlations (total τ = .19*; written τ = .06; oral τ = .09) indicate a lack of agreement between the two assessments around the cut-off point.

The overall correlation could be taken as support for the assumption of level equivalence, but all other analyses point towards a different conclusion. First, the descriptive statistics (see Table 4.2) indicate that the ITNA mean scores (standardized to a percentile scale for the purpose of comparison) are lower than those on STRT, both overall and for the separate components.

Table 4.2. Descriptive statistics (percentile scale)

	ITNA			STRT		
	Computer	Oral	Total	Written	Oral	Total
N	118	82	118	118	82	118
$\bar{X}$	59.31	50.97	48.06	66.76	73.16	67.68
SD	15.54	19.69	20.09	12.63	10.33	11.72
Med	58.42	48.75	51.63	67.33	72.54	69.19
SE	1.42	2.17	1.84	1.16	1.14	1.07

T-tests confirmed that the differences between STRT and ITNA are significant (total: t (236) = -9.20, p < 0.001; written: t (236) = -4.061, p < 0.001; oral: t (162) = -9.036, p < 0.001), with medium (written: d = -0.53) to large (total: d = -1.19; oral d = -1.41) effect sizes. The differences between the two tests remain significant within the interquartile range, with large effect sizes (total: W = 135, p < .000, d = -2.14; written: W = 775, p < .000, d = -1.04; oral: W = 190, p < .000, d = -1.16).

In order to determine whether the difference in mean scores meant that fewer candidates passed ITNA, a crosstab of STRT and ITNA pass/fail judgments was constructed. Table 4.3 confirms that more participants failed ITNA than STRT and that 24% of the population received a different pass/fail judgment. McNemar's binomial sign test shows that this difference is significant at p = .001. Additionally, the pass probability is significantly (W = 6010, p = .02) lower for ITNA (P_{ITNA}^{pass} = .35) than for STRT (P_{STRT}^{pass} = .50).

Table 4.3. Pass/Fail crosstab

		STRT		
		Fail	Pass	Total
ITNA	Fail	53	23	76
	Pass	5	37	42
	Total	58	60	118

Construct equivalence

In order to explain why more candidates failed ITNA than STRT, the written and the oral test components were examined.

Writing component

The first step in explaining the nature of the relationship between the written components was constructing a multiple linear regression model (table 4.4) to determine to what extent STRT task scores predicted the total ITNA score. The model explains 48% of the ITNA score variance. The strongest predictors are the writing-from-listening summarization task and the writing-from-reading argumentation task.

Table 4.4. ITNA computer scores as a function of STRT written scores

	STRT				
	ArgAudio	SummAudio	ArgRead	SummRead	Constant
B	.373	2.514	3.784	1.845	2.996
SE *B*	1.034	.781	1.155	.828	7.491
β	.031	.322	.280	.206	
p-value	*ns*	**	***	*	*ns*

Note. Total R^2 adjusted is .482 (p = < .001).

If all tasks are weighted equally, the MFRA reliably (.90) shows that $ITNA_{computer}$ is more difficult than $STRT_{written}$, mainly due to the relatively high difficulty of ITNA's vocabulary and grammar items as well as the relatively low difficulty of the argumentative STRT tasks. The MFRA model reliably (.93) identified three distinct levels of difficulty in the tasks: the borders between those levels are indicated with a dashed line in table 4.5. The most difficult tasks are the ITNA vocabulary and grammar tasks, and the least challenging ones are the STRT argumentative tasks. The dictation task in ITNA misfits the Rasch model and STRT's writing-from-reading argumentation task (STRT) is redundant.

Table 4.5. MFRA written tasks (equal weights)

Task	Test	Measure	SE	Infit MnSq
Language-in-use: Vocabulary (cloze)	ITNA	.43	.10	1.42
Language-in-use: Cloze (small)	ITNA	.40	.07	.71
Language-in-use: Vocabulary (MC)	ITNA	.38	.07	1.22
Language-in-use: Grammar (gaps)	ITNA	.33	.07	.72
Language-in-use: cloze (big)	ITNA	.10	.07	.97
Note-taking (listening)	STRT	-.01	.08	.88
Written summary (reading)	STRT	-.04	.08	.95
Structuring (drag-drop)	ITNA	-.07	.08	1.06
Multiple choice reading	ITNA	-.08	.09	1.40
Multiple choice listening	ITNA	-.15	.08	1.01
Dictation (fill in the gaps)	ITNA	-.18	.09	1.71
Argumentative writing-from-reading	STRT	-.51	.10	.42
Argumentative writing-from-listening	STRT	-.59	.11	.83

Summary statistics:

Candidate: Model, Random (normal): X^2 (*115*) = 98.4, p = .87

Task: Model, Fixed (all same): X^2 (12) = 164.5, p = .00

Speaking component

In order to investigate the similarities and differences between STRT and ITNA, a multiple linear regression analysis was run in which all STRT criteria were regressed onto the ITNA scores. The regression model explains 28% of the ITNA score variance (p < .001). The low adjusted R^2 (.28) hints at a possible discrepancy in the rating of STRT and ITNA. When highly similar tasks and criteria are used to score the same pool of candidates, one would expect the regression model to be a better fit. Consequently, in order to determine to what extent corresponding criteria fit the same underlying constructs, a PCA (see Table 4.6) was conducted on the standardized z-scores of the oral test components.

Table 4.6. Promax-rotated factor loadings

	TC 1	TC 2	TC 3
STRT			
Vocabulary	.92	…	…
Cohesion	.85	…	…
Grammar	.80	…	…
Fluency	.70	…	…
Pronunciation	.68	…	…
ITNA			
Vocabulary	…	.97	…
Fluency	…	.81	…
Grammar	…	.75	…
Cohesion	…	.58	…
Pronunciation	…	…	.96
Eigenvalue	3.32	2.54	1.16
Proportion explained	.47	.36	.17

Note. Factor loadings ≤ .3 were omitted.

The PCA confirmed that the ratings of corresponding criteria do not match; corresponding criteria in STRT and ITNA do not load onto the same factors, which one would expect if they measured the same underlying constructs. Instead, all STRT criteria cluster together, as do all ITNA criteria, except for *Pronunciation*.

Finally, the first Rasch model – in which the criteria were weighted equally in order to allow for a clear comparison – reliably (.98) identifies seven distinct difficulty levels in the STRT and ITNA criteria. In this model, the oral components of both tests cannot be reliably separated in terms of difficulty (reliability .00; X^2 (1) = .2, *ns*), but the measures of the criteria reveal some telling mismatches. Table 4.7 shows that *Pronunciation* in ITNA ranks as markedly difficult, and *Content* in STRT's argumentation task is disproportionately easy. The MFRA further shows that, except for *Vocabulary*, corresponding STRT and ITNA criteria invariably belong to a different difficulty band. This would most likely not be the case if both tests interpreted corresponding criteria in the same way. As such, the first MFRA confirms the PCA and adds an extra layer of

information to it; corresponding criteria probably measure different constructs, and – except for *Vocabulary* – do so at distinctly different difficulty levels.

Table 4.7. MFRA oral criteria (equal weights)

Criterion	Test	Measure	SE	Infit MnSq
Pronunciation	ITNA	1.34	0.20	1.49
Fluency	STRT	1.09	0.20	1.06
Coherence	STRT	0.65	0.20	1.13
Grammar	STRT	0.57	0.20	0.92
Content presentation	STRT	0.40	0.21	1.02
Pronunciation	STRT	0.23	0.21	0.97
Vocabulary	ITNA	-0.35	0.22	0.91
Vocabulary	STRT	-0.35	0.22	0.97
Initiative	STRT	-0.40	0.22	0.70
Coherence	ITNA	-0.89	0.23	0.80
Fluency	ITNA	-0.94	0.23	1.04
Grammar	ITNA	-1.37	0.23	0.45
Register	STRT	-1.69	0.24	0.80
Content argumentation	STRT	-5.52	0.39	1.85

Summary statistics:

Candidate: Model, Random (normal): $X^2(71) = 63.6$, $p = .72$

Task: Model, Fixed (all same): $X^2(13) = 419.1$, $p = .00$

The MFRA, with equally weighted criteria, offers a clear picture of the relative difficulty of the criteria, but in reality, not all criteria are weighted equally. STRT weighs linguistic criteria double to compensate for the amount of content criteria, and ITNA – having no content criteria – assigns a weight of 2 and 1.6 to *Grammar* and *Vocabulary*. A Rasch model that uses the actual weights of the criteria reliably (.88) identifies ITNA as the most difficult by half a logit. Applying actual weights instead of equal ones also changes the relative order of the criteria (see Table 4.8). All corresponding criteria now appear in distinctly different difficulty bands; *Grammar* in ITNA becomes the most difficult criterion by nearly two logits, and *Content* in the STRT argumentation task is roughly four logits easier than the second easiest criterion, misfitting the Rasch model, as does *Pronunciation* in ITNA.

Table 4.8. MFRA oral criteria (actual weights)

Criterion	Test	Measure	SE	Infit MnSq
Grammar	ITNA	2.89	0.33	1.11
Fluency	STRT	0.90	0.14	1.01
Vocabulary	ITNA	0.79	0.21	1.25
Pronunciation	ITNA	0.65	0.15	1.64
Coherence	STRT	0.47	0.14	1.04
Grammar	STRT	0.43	0.14	0.79
Pronunciation	STRT	0.02	0.15	0.88
Content (presentation task)	STRT	-0.13	0.21	1.06
Vocabulary	STRT	-0.46	0.15	0.81
Initiative	STRT	-0.55	0.15	0.76
Coherence	ITNA	-1.69	0.16	0.75
Fluency	ITNA	-1.74	0.16	1.01
Register	STRT	-1.76	0.16	0.79
Content (argumentation task)	STRT	-5.71	0.44	1.64

Summary statistics:

Candidate: Model, Random (normal): X^2 (70) = 66.1, p = .61

Task: Model, Fixed (all same): X^2 (13) = 679.3, p = .00

DISCUSSION

Flemish universities that accept STRT and ITNA certificates as equivalent measures of B2 proficiency make a claim about level equivalence. Validating such a high-stakes claim requires strong empirical evidence (Kane, 2013), which had not been presented to date. Flanders is not exceptional in this sense; in many different countries, university entrance policies contain similar claims of test equivalence, often without substantiation (McNamara & Ryan, 2011). For that reason, one of the wider aims of this study was to highlight the importance of empirically examining unfounded claims of level equivalence.

This study showed the correlation (r = .77**) between overall STRT and ITNA scores to be rather strong and comparable to coefficients reported in previous research (e.g., ETS 2010 (r = .73); Zheng & De Jong, 2011 (r = .75)). However, since overall correlations can be misleading (Lissitz & Samuelsen,

2007), supplementary analyses were conducted, all of which revealed that STRT and ITNA do not map onto each other quite as seamlessly.

When considering only the interquartile range, the correlation between STRT and ITNA scores becomes virtually zero, indicating that test takers without a distinctly strong or weak profile were assessed differently by STRT and ITNA. This hypothesis is confirmed by an analysis of the descriptive data, which shows significantly lower mean scores on ITNA than on STRT. Perhaps the most telling piece of evidence to discount the claim of level equivalence is the significant difference in pass-fail judgments; 24% of the participants received a different outcome on STRT and ITNA. In most cases of disagreement, candidates failed ITNA but passed STRT. Accordingly, the probability of passing STRT is significantly higher than the probability of passing ITNA.

The evidence presented in this study casts doubt on any claim of level equivalence. Apart from the overall correlation, none of the analyses offered evidence in support of the policy claim, which, as a result, cannot be considered valid (Kane, 2013). Moreover, following Phillips's decision rule (Kane, 2013; Phillips, 2007), the university entrance policy is unlikely to solve the problem it was intended to fix, that is, assuring a consistent minimum language level among the international student population.

These results have implications beyond the immediate research context. First, they reaffirm the danger of assuming that different tests linked to the same level of the CEFR are equally difficult (Green, 2017). Policy makers often rely on these levels when determining language requirements (Fulcher, 2012b), but since CEFR levels are broad and leave room for interpretation (Alderson, 2007; Fulcher, 2004), direct cross-test comparison of the kind that was presented in this study might be a safer, more robust option. This study indicates that even tests which have been linked to the same CEFR level may differ substantially in pass/fail judgments. Secondly, this study confirms the criticism raised against using correlational data in validation research (Lissitz & Samuelsen, 2007; Norris, 2016), and underscores the importance of supplementing purely correlational results with impact data (i.e., pass/fail decisions) and with information concerning the nature of a relationship between two variables (i.e., construct equivalence).

The construct equivalence analyses indicate substantial differences between comparable STRT and ITNA tasks. STRT task scores explain 48% of the score variance on ITNA's written component, and STRT's argumentative tasks contribute less to the regression model than the summarization tasks. This is not entirely surprising, because argumentative tasks require knowledge transformation, whereas summarization tasks rely on repetition and are more in line with the multiple-choice items found in ITNA. The MFRA identifies the argumentative STRT tasks as the easiest tasks in both tests, which could explain why they contribute little to the regression. This observation, combined with the fact that ITNA's vocabulary and grammar tasks are nearly half a logit more

difficult than the most difficult STRT task, explains why ITNA's written component is the most difficult.

Even though the oral components of STRT and ITNA are highly similar in terms of task types and rating criteria, the evidence in favor of construct equivalence is weak. In the multiple linear regression analysis, STRT criteria explained just 28% of the ITNA score variance, the PCA showed that corresponding criteria from both tests do not load onto the same factor, and the MFRA with equally weighed criteria mapped most of the corresponding criteria into different difficulty bands. The analyses all indicate that corresponding criteria likely measure different constructs. Moreover, an MFRA, using the actual weights of the criteria, reliably shows that the oral component of ITNA is the hardest because it assigns the greatest weight to comparatively difficult criteria (*Grammar* and *Vocabulary*), and because a large proportion of the STRT scores is derived from relatively easy content criteria. Importantly, the difference between STRT and ITNA only becomes substantial when the actual weights are operationalized in the Rasch model. A Rasch model with equal weights for every criterion did not reliably differentiate between the difficulty levels of the two tests.

It is important to note that ITNA's greater level of difficulty does not automatically imply that it is a better university entrance language test. Answering that question requires a different set of data. The data used in this dissertation do provide substantial evidence to argue that STRT and ITNA measure different constructs, even when the tasks and the criteria are highly similar. Additionally, the analyses show that the relative importance assigned to content, grammar, and vocabulary in STRT and ITNA is the most likely cause of the differing difficulty levels in the oral and the written components.

In terms of justice, the situation in Flanders can be improved, perhaps even without invoking drastic measures. It may suffice to extend to test takers the same service that customers of other paid services receive. Before purchasing a service, people typically request and receive detailed information about it, allowing them to make a balanced choice. If the information they received prior to purchase was misleading, customers have the right to complain and revoke the contract. The same could apply when people select a high-stakes test based on inaccurate information or unfounded assumptions. The need "*to provide all potential test takers with adequate information about the purposes of the test, the construct (or constructs) the test is attempting to measure and the extent to which that has been achieved*" (ILTA, 2007, p. 3, emphasis added) is not a new concern, but it is one that deserves renewed attention. It would benefit the justice of testing policies if test takers received accurate information about the actual differences and similarities between tests that are presented as equivalent options.

CONCLUSION: ASSUMPTION 3

The data presented in this chapter do not indicate that the STRT and ITNA scores can be considered equivalent. Roughly one in four participants received a different pass/fail score. ITNA appears to be the harder test, largely because of the language-in-use tasks, and because of the role vocabulary and grammar play in the test construct. The next chapter zooms in on equivalence of corresponding rating criteria in the oral test components.

CHAPTER 4
CRITERION EQUIVALENCE

This fourth chapter focuses specifically on equivalence of the STRT and ITNA components that are most comparable. It examines how corresponding CEFR-based criteria that are operationalized in highly comparable speaking tasks map onto each other.

Prior to 1800, when Henry Maudslay developed the first standardized screw thread, nuts and bolts were not easily interchangeable. Maudslay introduced a common standard, and changed the life of every plumber to this day. Standards help to achieve transparency, uniformity and interchangeability – at least in hardware. Language performance, in all its idiosyncratic and contextual variation, does not easily permit such standardization. Nevertheless, fair and valid testing hinges upon score comparability and score transparency (Kane, 2013). The first concept implies that scores on tests that target the same audience and share the same purpose can be meaningfully compared. The second concept – score transparency – entails that test scores have clear meaning to users. Candidates and admission officers need to be able to meaningfully interpret test scores (Alderson, 1991), and raters need to have the same conception of the same level. This makes it all the more striking that, in language testing standards are constant in one sense: they vary.

Currently, numerous language performance standards exist alongside each other. The ACTFL standards resulted from national efforts (ACTFL, 2012), the STANAG 6001 standards are used within supranational organizations (NATO, 2014), and still other standards have been developed by testing organizations in the form of rating scales. Since different organizations use different scales, it is not easy for test takers or test users to interpret scores and compare them with other tests (Gomez, Noah, Schedl, Wright, & Yolkut, 2007). In Messick's (1989) approach to validity, which considers score use essential to a validity argument, a lack of score transparency presents a problem, which could potentially be resolved if all tests were linked to the same universally accepted levels and standards of performance. In Europe and beyond, the CEFR (Council of Europe, 2001) is widely considered as such a standard (see Chapter 1). The CEFR's uptake has been widespread, but it has not been empirically validated in every intended or unintended context of use.

This chapter examines the use and usefulness of the CEFR in the context of rating scale design. More specifically, it contributes evidence regarding Assumption 3 by exploring to what extent the same CEFR descriptors have been

similarly operationalized and interpreted in the oral components of two university entrance language tests.

CRITERION EQUIVALENCE AND THE CEFR

Since its publication in 2001, the CEFR (Council of Europe, 2001) has become widely used and adopted by test developers, policy makers, teachers, publishers and candidates alike. It has come to be seen as a common currency in language performance levels (Figueras, 2012), and in Europe it is now the leading framework in language testing (Figueras, 2012; Little, 2007; Papageorgiou, Xi, Morgan, & So, 2015). The CEFR is so influential that it has become necessary for tests to link to it in order to gain recognition within Europe (see Chapter 1). Outside of Europe too, many scoring systems and performance standards have been mapped onto the CEFR (e.g., Bärenfänger & Tschirner, 2012; Baztán, 2008; Tschirner, Bärenfänger, & Wisniewski, 2015 for ACTFL; Tannenbaum & Wylie, 2008 for TOEFL iBT; Swender, 2010 for STANAG 6001; Zheng & De Jong, 2011 for PTE Academic; also see Green, this issue). Theoretically, a framework that has received such wide recognition by all parties involved could address the score transparency and comparability concerns Kane (2013) and Alderson (1991) raised. In practice however, there are issues.

Even though the goals of the CEFR in its current form are descriptive, not normative (North, 2014a), achieving score comparability across tests was one of the primary goals of its earliest drafts (van Ek, 1975: 8). Today too, in many European contexts, the CEFR level descriptors are used in a normative way, as performance standards, or as labels to facilitate score transparency (Roever & McNamara, 2006; O'Sullivan & Weir, 2011; Fulcher, 2012). With score transparency in mind, many test developers are using CEFR descriptors as the basis for rating scale development, but even though treating the CEFR as a heuristic is common practice (Weir, 2005b; North, 2014a; 2014b), it is not unproblematic. First, the CEFR offers guidance on essential test development matters, such as test purpose, response format, time constraints, and topic (Weir, 2005b). Two tests could have the same CEFR level, but very different specifications, and it would be wrong to consider them equivalent simply because they share a CEFR label (Green, 2017; Taylor, 2004). Secondly, because the CEFR is context and language-independent, test developers need to add specific details to the descriptors when using it in a rating context (Harsch & Martin, 2012). This necessary step may cause two tests to deviate in their interpretation of the CEFR levels, resulting in reduced comparability. In fact, CEFR descriptors have been criticized for their vagueness and inconsistencies, both within and across levels (Alderson, 2007; Harsch & Rupp, 2011; Papageorgiou, 2010) and may suffer from "descriptional inadequacy" (Fulcher, Davidson, & Kemp, 2011: 8), leaving room for

dissimilar interpretations. Since there is ample evidence that even trained raters interpret the same test-specific criteria differently (Deygers & Van Gorp, 2015; Lumley, 2002; 2005) and that also trained raters' experience and background may influence the score that is assigned (Barkaoui, 2011), there can be no guarantee that different test developers interpret the same CEFR descriptors in the same way. To the best of our knowledge, no study has yet compared the ratings of two high-stakes tests using corresponding CEFR-based criteria.

Nonetheless, quite a few studies have discussed rating scale construction in relationship to the CEFR (Galaczi, ffrench, Hubbard, & Green, 2011; Harsch & Martin, 2012; Papageorgiou, 2015). These studies typically discuss fitting CEFR descriptors to rating scale logic by rectifying descriptor vagueness and by straightening blurred lines between levels (Alderson, 2007; Papageorgiou, 2010). In addition, Galaczi et al. (2011) have highlighted the positive wording of CEFR descriptors and the brevity of certain CEFR scales as matters of ongoing concern during rating scale construction and rater training. Deygers & Van Gorp (2015) showed that a CEFR-based rating scale that was iteratively constructed together with raters did not guarantee a uniform interpretation of the descriptors, in spite of high inter-rater reliability indices. In this regard, Harsch and Rupp (2011) rightfully stressed the need for a high level of analytic detail in CEFR-based scales in order to compensate for the broadness of the initial descriptors.

The abovementioned studies show how individual test developers have operationalized CEFR descriptors in their rating scales to fit the purpose of a test. Other documents describe how some tests have been aligned with the CEFR (e.g., Tannenbaum & Wylie, 2008; Khalifa & ffrench, 2009; De Jong, Becker, Bolt, & Goodman, 2011 for TOEFL iBT, IELTS and Pearson PTE respectively). Green (2017) has scrutinized some of these reports in an effort to understand the varying ways in which the major English tests have established their CEFR links. He uncovered that the linking methodologies used by these high-stakes university entrance language tests diverged so substantially that it would be misguided to presume comparability of the B2 levels.

To date, little if any CEFR research has been comparative. Concurrent analyses themselves are not new to the language testing endeavor however, and have actually been central to test validation (Chapter 3 contains an overview). Nevertheless, no studies, concurrent or otherwise, have analyzed empirical data to compare the interpretation and operationalization of CEFR descriptors across tests. Nevertheless, it could be argued that exactly this kind of research determines whether the CEFR can act as a catalyst for increased score transparency and score comparability, which was one of its original goals (Van Ek, 1975). The current study thus addresses an important gap in the literature by examining the potential of the CEFR in facilitating score transparency in tests that share the same purpose, the same population and employ corresponding rating criteria that are based on the same CEFR descriptors.

RESEARCH QUESTION

The study described in this chapter examines to what extent corresponding STRT and ITNA criteria lead to the same scores for the same candidates in the same way.

RQ *Do the two tests apply the same CEFR-based level descriptors in the same way for similar task types?*

In order to systematically answer this RQ, four sub questions were identified:

a. How much do the STRT and ITNA criteria deviate from the original CEFR descriptors?
b. Can corresponding CEFR-based levels in both tests be considered truly equivalent?
c. Are corresponding CEFR-based criteria likely to measure the same construct?
d. Are corresponding CEFR-based criteria equally difficult in both tests?

If both tests interpret the same CEFR descriptors in the same way, high overall correlations should carry through down to the criterion level. If the same candidates are rated highly dissimilarly on corresponding criteria, using the CEFR as a standard for score transparency may not be warranted, since it may create a false sense of uniformity.

PARTICIPANTS & METHODOLOGY

The analyses in this chapter are based on the scored performances of 82 $L2_F$ participants who took the oral components of STRT and ITNA within the same week (see Chapter 3).

STRT & ITNA rating scales

Typically, the oral STRT or ITNA components do not take more than 25 minutes, including preparation time (see Appendix 1 and 2). In both tests, candidates interact with a trained examiner during the oral component, which consists of a presentation and an argumentation task. The argumentation task invites the candidates to weigh a number of alternative solutions to a problem, and to argue why their choice is the better one.

Five oral rating criteria are included in both tests: *Vocabulary*, *Grammar*, *Coherence*, *Pronunciation*, and *Fluency*. For scoring these criteria, both tests

employ analytic band descriptors that are based on the A2, B1, B2, and C1 levels in the corresponding CEFR scales. In both tests the cut off level for each criterion is B2, except for *Pronunciation* and *Grammar*, where the ITNA uses B1 and B2+ respectively. Both tests developed their rating scales by drawing on the original descriptors, but both made choices based on their interpretations of the CEFR descriptors. ITNA rating scale designers often copied the original CEFR text and supplemented it with language-specific examples, identifying typical errors of users at a given level. The STRT criterion descriptors deviated from the original wording more often, in order to make the original descriptors more concrete and easier to grasp for the novice raters they often employ (see Deygers & Van Gorp, 2015).

The oral ITNA performances are scored immediately after the test by two trained raters who come to one composite score for each of the five criteria. ITNA examiners and raters tend to be experienced L2 teachers of Dutch who typically attend training at least once a year and score oral tests at different times throughout the year. The STRT performances are recorded, and subsequently two independent trained raters – who are usually novice raters with a background in linguistics or communication – separately score each task. They receive a two-day training, and take part in a trial rating session to establish their consistency and reliability.

Data analysis

Determining whether both tests apply the same CEFR-based rating criteria in the same way required recoding certain scoring categories. The STRT rating scale distinguished four proficiency levels (A2, B1, B2, and C1), but the ITNA had six or seven, since it includes the plus levels of B1, B2 and occasionally A2. After consulting with the ITNA coordinators, these double bands were merged in order to come to a four-band scale, which facilitates direct comparisons across scales. Below, the analyses are discussed by subquestion:

How much do the STRT and ITNA criteria deviate from the original CEFR descriptors?

The STRT and ITNA rating criteria were compared to each other and to the Dutch translation of the CEFR, using the Jaccard similarity index. The index provides a very simple quantification of the similarity between descriptors. It has been applied in different forms in the field of information retrieval and text comparison (Manning, Raghavan, & Schütze, 2009). The Jaccard index expresses the similarity between two descriptions as their overlap in terms, more specifically the ratio of the number of unique terms present in both texts and the number of unique terms in either of the texts. The Jaccard index becomes 1 if

both texts use the exact same set of words, independent of how often these terms are repeated, and it decreases when the terms used in both texts diverge.

In order to make the comparison of the descriptors used in this study more robust, the texts were automatically pre-processed using a standard stemming algorithm for Dutch. This means that all words were stemmed (e.g., plural endings removed) and all non-informative words (such as "of" and "with", as defined by the Python NLTK Dutch stopword list) were removed.

Can corresponding CEFR-based levels in both tests be considered truly equivalent?

In order to determine the equivalence of the same levels in corresponding criteria, frequency distributions of CEFR-based scores were supplemented with probability estimates of attaining the B2 level. For every criterion the probability of attaining a score of B2 or higher was estimated. The strength of the relationship between corresponding criteria was calculated using Kendall's Tau. To determine the level of agreement between the two tests, linear weighted kappa (K_w) was used. K_w is a variation on Cohen's kappa, which measures the level of agreement between ordinal data sets (Sim & Wright, 2005), whereby 0 indicates no agreement except one stemming from chance, and values above .8 can be read as almost perfect agreement (Landis & Koch, 1977; Vanbelle & Albert, 2009). Usually, weighted kappa is used to determine rater agreement, but in this study it served as an additional metric to determine whether the STRT and ITNA raters scored the same candidates in the same way for corresponding criteria.

Are corresponding CEFR-based criteria likely to measure the same construct?

The ITNA rating scale consists of five criteria: *Vocabulary*, *Grammar*, *Coherence*, *Pronunciation*, and *Fluency*. These criteria also occur in the STRT rating scale, in addition to others (e.g., *Content*, *Register*). If the STRT and ITNA raters interpreted the same criteria in the same way, the STRT scores on the shared criteria would explain a large proportion of the total ITNA score variance. In order to investigate this, three multivariate linear regression models were constructed. Each model took the following general form:

$$\text{ITNA}_{\text{Total}} = (b_0 + b_1 \text{criterion}_{1i} + b_2 \text{criterion}_{2i} + \ldots + b_n \text{criterion}_{ni}) + \varepsilon_i$$

Three regression models were run, and compared in terms of R^2 using an Anova. The first regression model included the five criteria from the two STRT tasks that significantly correlated with the ITNA criteria at $\tau > .3$. The second model included the seven criteria that significantly correlated, regardless of the strength of the correlation. The final model included all the STRT criteria. Prior to running the regression analyses, the assumptions were checked: The proportion

of cases with large residuals was acceptable (4% in the oral component after removal of two outliers), Cook's distance was <1, no cases were larger than three times the average leverage, the covariance ratio was satisfactory, and the multicollinearity and independence assumptions were supported (Norris 2015; Purpura, Brown, & Schoonen 2015).

Next, to determine the relationship between individual criteria, a linear regression model was constructed for every ITNA criterion as a function of the same STRT criterion.

Are corresponding CEFR-based criteria equally difficult in both tests?

In a multifaceted Rasch (MFRA) measurement analysis, a test score is seen as the result of an interaction between different facets, such as test-taker ability, task difficulty, rater severity and criterion difficulty (McNamara, 1996). In MFRA the effect of all these variables on the score is taken into consideration and mapped onto the same logit scale. In this study, all comparable STRT and ITNA ratings for the same candidates were combined in the same MFRA model and the Facets program was used to estimate criterion difficulty. Of interest are the difficulty measures (a higher measure indicates a more difficult criterion), the strata index (which shows whether different measures also translate into different levels of difficulty that can be separated reliably) and the fit statistics (InfitMnSq). The closer the value of these fit statistics is to 1, the better the observed data fit the Rasch model. A criterion that has fit statistics in the range between .50 and 1.5 is considered to have an acceptable model fit. Lower values indicate overfit (i.e., redundancy) and higher values indicate misfit (Linacre, 2012; Barkaoui, 2014).

The analyses were conducted using *R* (*psych*, *irr*, *Hmisc*, *QuantPsyc*, *car*, and *ggplot2* packages), Facets (Linacre, 2015), and Python (with the NLTK library).

RESULTS

The Jaccard index (see Table 5.1) shows that on the whole the wording of the ITNA criteria stays closer to the exact wording of the CEFR than the wording of the STRT criteria. For example, the lowest Jaccard index for *ITNA ~ CEFR* is .27, but three out of five *ITNA ~ CEFR* indices are substantially lower than that ($J \leq$ .10). Since the wording of the descriptors in both tests deviates substantially from the CEFR original, it is logical that the overlap between the STRT and ITNA descriptors is typically not too big. The rating scale descriptors of *Pronunciation* in both tests stay closest to the CEFR wording, as a result of which the overlap between the STRT and ITNA descriptors is the largest for this criterion (J = .44).

Table 5.1. Jaccard index for rating descriptor pairs

	ITNA ~ CEFR	STRT ~ CEFR	ITNA ~ STRT
Vocabulary	.53	.15	.10
Grammar	.30	.10	.06
Coherence	.27	.09	.08
Pronunciation	.80	.40	.44
Fluency	.88	.26	.29

For reasons of confidentiality, the full rating scale descriptors cannot be repeated in their entirety, but a few examples, taken from confidential STRT and ITNA rating scale documents, may serve to illustrate how CEFR descriptors were paraphrased.

The B2 criterion for *Vocabulary* in ITNA adds to the CEFR descriptor: "has a good range of vocabulary for matters connected to his/her field and most general topics *and does not only use high-frequency words*" (my italics). The STRT vocabulary criterion, on the other hand, reads: "the lexical variation in the performance is sufficient to prevent frequent repetition of words". In both tests, the descriptors for *Coherence* include additions to the CEFR wording. The ITNA focuses on the sentence level: "sentences are linked logically and appropriate connectors are used when required". The STRT raters are required to consider the text level as well: "The performance is one coherent whole (…) connectors are mostly used correctly and support the overall coherence". *Pronunciation* in STRT repeats the original B2 descriptor, but it is supplemented with a B1 characteristic: "The pronunciation is clear and natural, *but with a foreign accent*" (my italics). This addition does not occur in the ITNA rating scale, where the cut off point for *Pronunciation* is B1, not B2. The *Fluency* descriptor in the ITNA rating scale has been literally copied from the CEFR: "can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions, there are few noticeably long pauses." A few additions were made in the STRT descriptor: "can produce stretches of language with an even tempo; although he/she can be hesitant as he/she searches *for the right expression*, there are *few noticeable or distracting* pauses" (my italics).

After examining the correspondence in the wording of the rating criteria, the scores were analyzed. It was determined whether the same levels of corresponding criteria in the STRT and ITNA rating scales can be considered equivalent. Table 5.2 shows how often CEFR levels were assigned to the same criteria in the ITNA and in both STRT tasks (coherence is not a rating criterion in the STRT argumentation task). In most cases the mode corresponds with the B2 level. In other words, on most criteria, most candidates scored B2 (e.g. 53 ITNA test takers scored B2 on *Vocabulary*, and 8 scored C1). For $Grammar_{STRTpres}$ and

Pronunciation$_{ITNA}$, B1 was the level most often assigned. No candidate scored A2 for *Grammar*$_{ITNA}$ or *Coherence*$_{ITNA}$ (ITNA assigned 0 A2, and 16 B1 ratings on *Coherence*, while STRT scored 9 performances A2 and 31 B1).

Table 5.2. Frequencies of assigned CEFR levels

		A2	B1	B2	C1
Vocabulary	ITNA	5	16	53	8
	$STRT_{arg}$	3	17	40	22
	$STRT_{pres}$	7	20	40	15
Grammar	ITNA	0	4	67	11
	$STRT_{arg}$	4	24	44	10
	$STRT_{pres}$	5	34	33	10
Coherence	ITNA	0	16	49	17
	$STRT_{pres}$	9	31	31	11
Pronunciation	ITNA	11	40	23	8
	$STRT_{arg}$	2	24	45	11
	$STRT_{pres}$	3	32	37	10
Fluency	ITNA	1	17	50	14
	$STRT_{arg}$	5	27	36	14
	$STRT_{pres}$	12	30	37	3

For most criteria the probability of any given candidate to attain a score of B2 or more was found to be higher on ITNA than on either of the STRT tasks (see Table 5.3).

Table 5.3. Probability of attaining B2 or higher on STRT and ITNA

	$P_{B2}^{STRTarg}$	$p^{\#}$	P_{B2}^{ITNA}	$p^{\#\#}$	$P_{B2}^{STRTpres}$
Vocabulary	.76	1.0	.74	.362	.67
Grammar	.66	.000	.95	.000	.52
Coherence			.80	.000	.51
Pronunciation	.68	.000	.38	.009	.57
Fluency	.61	.013	.78	.000	.49

Note. $p^{\#}$: p-value for the difference in probability between $P_{B2}^{STRTarg}$ and P_{B2}^{ITNA}

$p^{\#\#}$: p-value for the difference in probability between $P_{B2}^{STRTpres}$ and P_{B2}^{ITNA}

The *p*-values in Table 5.3 refer to the difference in probability between the two STRT tasks and the ITNA results. A given candidate has a 38% probability of attaining a B2 pronunciation score on ITNA. The same candidate may, however, have a 68% probability of being rated B2 on the same criterion if he or she performs the STRT argumentation task. The difference between these probabilities is significant. In fact, vocabulary excepted, there is a consistent

significant difference between the probability of attaining a score of at least B2 on the ITNA tasks or one of the STRT tasks ($p < .05$). This indicates that the B2 threshold is interpreted or operationalized differently on the STRT and on the ITNA test.

The frequencies in Table 5.2 show regular discrepancies in STRT and ITNA judgments and the probabilities in Table 5.3 indicate that STRT and ITNA judgments differ in severity from one criterion to the next (e.g., a B2 score on coherence is significantly harder to reach on STRT than on ITNA). As such, there likely is a different distribution in the ITNA and STRT scores for corresponding criteria.

Table 5.4 shows the ITNA score on both tasks combined in relation to the STRT argumentation task and the STRT presentation task. The relationship between the corresponding STRT and ITNA criteria is generally medium to low ($\tau \leq .39^{**}$) and the agreement is generally weak ($k_w \leq .22$). This implies that corresponding STRT and ITNA criteria might not map onto each other well. The relationship is weakest for *Vocabulary* and *Pronunciation*. The correlation between the sums of these five corresponding criteria is moderate as well ($\tau = .37^{**}$).

Table 5.4. Relationship between corresponding STRT and ITNA criteria

	ITNA ~ $STRT_{arg}$		ITNA ~ $STRT_{pres}$	
	τ	k_w	τ	k_w
Vocabulary	.153	.031	$.212^{*}$	.091
Grammar	$.336^{*}$	$.208^{***}$	$.351^{**}$	$.184^{***}$
Coherence			$.386^{**}$	$.216^{***}$
Pronunciation	.117	$.122^{*}$	$.212^{*}$	$.207^{**}$
Fluency	$.336^{**}$	$.215^{**}$	$.315^{**}$	$.134^{*}$

Note. * = $p < .05$, ** = $p < .01$, *** = $p < .001$
Overall correlation for summed criteria $\tau = .37^{**}$

Having determined that the same levels in corresponding criteria are unlikely to be equivalent, multivariate linear regression was used to determine to what extent scores on the criteria that the STRT shares with the ITNA, predicted the ITNA scores. Three models were run. The first included the five STRT criteria from the two STRT tasks which significantly correlated at $\tau > .3$. This model explained 26% of the ITNA score variance ($R^2_{adj} = .2585$, $p < .000$). The second model included the seven STRT criteria that correlated significantly with the corresponding ITNA criterion, regardless of the strength of the relationship. The second model accounted for 27% of the total ITNA score variance ($R^2_{adj} = .2706$, $p < .000$), but did not significantly improve the model fit of the data, in comparison with the first model ($F(2, 74) = 1.6334$, $p < .000$). The third multivariate linear regression model included all nine STRT criteria that had corresponding ITNA

criteria, and explained 26% of the ITNA score variance (R^2_{adj} = .2603, p < .001). Since this model was not a significantly better predictor than the first ($F(4, 72)$ = 1.0476, p < .39) or the second ($F(2, 72)$ = 0.4828, p < .62), Table 5.5 shows the regression results of the first model. It indicates that only one predictor (*Fluency*, in the argumentation task) significantly contributes to the model. Given the sample size, the estimates of the individual predictors should be treated with some caution, but generalizing from the overall model fit can be done with a degree of confidence (Field, Miles, & Field, 2012). All things considered, the multiple regression analyses show that no more than 27% of the ITNA score variance can be explained by scores on corresponding STRT criteria.

Table 5.5. Multivariate linear regression: $ITNA_{total}$ ~ $STRT_{arg+pres}$

	B	*SE B*	*β*	*p*
(Constant)	.457	3.811		.905
Grammar $STRT_{arg}$	1.547	1.851	.132	.406
Grammar $STRT_{pres}$	2.633	2.056	.241	.204
Fluency $STRT_{arg}$	3.709	1.436	.375	.012*
Fluency $STRT_{pres}$	-2.251	1.604	-.222	.165
Coherence $STRT_{pres}$	1.204	1.365	.128	.381

Note. Total R^2 adjusted is .2585 (p < .000).

When corresponding ITNA and STRT criteria were used in a pairwise linear regression (see Table 5.6), the same trend emerges. Regression models with statistical significance (p < .05) based on the STRT criteria never explained more than 17.2% of the score variance in corresponding ITNA criteria.

Table 5.6. Linear regression on criterion level: $ITNA_{total}$ ~ $STRT_{total}$

	r^2_{adj}		*B*	*SE B*	β	*p*
Vocabulary	.022					.156
		(Constant)	2.279	.336		.000
		$STRT_{arg}$	-.076	.211	-.081	.719
		$STRT_{pres}$	.252	.200	.281	.211
Grammar	.172					.000
		(Constant)	2.306	.189		.000
		$STRT_{arg}$	.122	.103	.196	.239
		$STRT_{pres}$	.156	.096	.267	.109
Coherence	.170					.000
		(Constant)	2.148	.215		.000
		$STRT_{pres}$	.325	.077	.425	.000
Pronunciation	.067					.017
		(Constant)	1.664	.400		.000
		$STRT_{arg}$	-.339	.250	-.276	.179
		$STRT_{pres}$	.605	.240	.513	.014
Fluency	.133					.001
		(Constant)	1.983	.261		.000
		$STRT_{arg}$	.213	.127	.259	.098
		$STRT_{pres}$	.135	.130	.160	.305

The results of the regression analyses above indicate that corresponding criteria are unlikely to measure the same construct at the same level. To ascertain whether the same criteria are equally difficult in the two tests, a multifaceted Rasch analysis was carried out, based on all corresponding rating criteria. This model (see Table 5.7) reliably showed that when only the five common criteria are used, STRT is the more difficult test. As the previous chapter showed, this does not imply that STRT is also the most difficult test overall.

Table 5.7. MFRA: STRT and ITNA, arranged by measure

	Measure	SE	InfitMnSq
STRT	.23	.09	1.02
ITNA	-.23	.10	.96

Note. Model, Sample: Separation 3.28, Strata 4.71, Reliability.92

Table 5.8 shows the results of the criteria measurement in the Rasch analysis. Importantly, these results showed that corresponding criteria were never included in the same difficulty bands. This implies that the difficulty level of every ITNA criterion is significantly different from its STRT counterpart. Moreover, the Rasch output generally aligns well with the probabilities displayed

in Table 5.3. For example, pronunciation in ITNA is the most difficult criterion in the Rasch table, and also had the lowest probability score. The probability scores for vocabulary were not significantly different, and in this table too, the measures of the vocabulary criteria of both tests are mapped closest to each other. Nevertheless, in spite of the results pertaining to the vocabulary scores, this study has yielded no data to indicate that corresponding CEFR-based criteria used to measure the same candidates in near-identical tasks can be considered equivalent.

Table 5.8. MFRA STRT and ITNA criteria, arranged by measure

Criterion	Test	Measure	SE	Infit MnSq
Pronunciation	ITNA	1.14	0.20	1.52
Fluency	STRT	0.47	0.20	1.14
Coherence	STRT	0.01	0.21	1.21
Grammar	STRT	-0.07	0.21	0.85
Pronunciation	STRT	-0.42	0.21	0.98
Vocabulary	ITNA	-0.6	0.22	0.94
Vocabulary	STRT	-1.02	0.22	0.88
Coherence	ITNA	-1.15	0.23	0.71
Fluency	ITNA	-1.21	0.23	0.99
Grammar	ITNA	-1.64	0.24	0.48

Summary statistics:

Candidate: Model, Random (normal): $X^2(71) = 63.6$, $p = .72$

Task: Model, Fixed (all same): $X^2(13) = 419.1$, $p = .00$

Importantly, the order of the corresponding criteria matches the order of Table 4.7 above.

DISCUSSION

Every analysis in this chapter confirms the trend observed in the previous chapter. There is little, if any, evidence to confirm an assumption of equivalence between STRT and ITNA. Corresponding CEFR-based criteria in the ITNA and STRT rating scales are not equivalent. If they were, the correlations would be stronger, the kappa values would show more agreement, the linear regression model would explain more variance and the same criteria would fall within the

same Rasch difficulty bands. One explanation for the divergences can be found in the rating scale descriptors. Even though both tests started from the same CEFR descriptors, they diverged in interpretation and operationalization. The Jaccard index indicated that the descriptors are indeed quite dissimilar, as seen in some of the operationalizations discussed above. In short, the statistical analyses in this study fail to confirm the assumption that the STRT and ITNA descriptors interpret the same CEFR levels in an equivalent way, and provide arguments to the contrary. Both tests have developed rating scales from the same source and adopted the same level system, but the relationship between equivalent criteria is weak. If the CEFR levels were true, unequivocal standards, this should not occur, but given the nature of the CEFR descriptors, the findings are not unexpected.

The root of the problem lies not so much in the CEFR itself as in the reification of its levels as standards (Fulcher, 2004). The CEFR is often referred to as a gold standard, because it is so eagerly used by all parties involved in European language testing, but there is one very important difference: exactness. The collective agreement that exactly one ounce of gold would trade for exactly $20.67 made the gold standard the backbone of the global economic system for decades. The CEFR intentionally lacks such exactness, however, which makes it unusable as a standard. CEFR levels do not exist outside of the minds of the practitioners, and B2 is not an entity. As a standard, it shares less resemblance to screw threads or monetary systems than to primary colours. The colour blue has a marked beginning and an end, but encompasses a range from light aquamarine to dark navy; it would be wrong to argue that only Pantone 2736C is the true blue. Likewise, it is problematic to consider the B2 level in one rating scale equivalent to the next, simply because both have been based on the same broad level.

CONCLUSION: ASSUMPTION 3

The results of Chapter 4 are in line with those of the previous chapter: There are no data to support the presumption that the level or the construct of STRT and ITNA are equivalent. The findings from this chapter show that STRT actually uses corresponding criteria in a stricter way. Nevertheless, ITNA is the hardest test due to the weighting of *Grammar* and *Vocabulary*, and due to the fact that STRT assigns relatively great importance to content criteria, which are comparatively easy.

CHAPTER 5
COMPARING L1 AND L2 PERFORMANCE

This chapter focuses on Assumption 4; the language proficiency level of first-year university students with a Flemish secondary school degree. Additionally, the study presented here explores to what extent Flemish and international L2 students perform differently on the same written STRT tasks, and whether the L2 learners who learned Dutch in Flanders or at their home institution perform differently on the same writing tasks.

Before moving on, a small note on terminology is needed. Within the group of Flemish students, a distinction is made between L1 users and Generation 1.5 students (G1.5). The term L1 user will be used to refer to a student whose first or home language is the same as the official language of instruction (Gorter & Cenoz, 2012), which in our case is Dutch. Students whose home language is different from Dutch, but who acquired Dutch during all or part of their schooling, will be called G1.5 students (di Gennaro, 2009, 2013, 2016; Harklau, Losey, & Siegal, 1999). Other research might refer to this group as sequential bilingual learners (Paradis, 2007; Pérez-Tattam et al., 2013). Under the current Flemish university entrance policy, both Flemish L1 users and G1.5 students can register for university studies without taking a B2 language test if they have graduated from a Dutch-medium secondary school.

Additionally, this chapter also distinguishes between two groups of international L2 students: those who studied Dutch at a Flemish language school ($L2_F$) prior to taking the university entrance language test, and those who studied Dutch at an international language school in their country of origin ($L2_I$).

RESEARCH INTO L1 AND L2 PERFORMANCE

Administering a language test as a gatekeeping instrument to international L2 students alone relies on the unsubstantiated assumption that students with a Flemish secondary school degree will meet the language demands that are required of international applicants. If this assumption is true, the university entrance policy justifiably exempts this group of students from additional language screening. If it were untrue however, and if not all students who are exempt from taking a test pass it, the university entrance policy could be considered to apply unequal standards to different populations (Hamilton, Lopes, McNamara, & Sheridan, 1993) – which could raise justice-related concerns.

Native speaker performance

Hulstijn posits that not all L1 users can be expected to perform writing tasks at the same level as some L2 learners (Hulstijn, 2015; Hulstijn, 2011) and proposes to differentiate between Basic Language Cognition (BLC) and Higher/Extended Language Cognition (HLC). BLC is restricted to the use of oral skills in common, everyday language situations. HLC involves lexically, syntactically and cognitively more complex language in both oral and written forms and includes "topics addressed in school and colleges" (Hulstijn, 2015: 22). While he does not explicitly differentiate between BLC and HLC in CEFR terms, it can be inferred that performing a writing task at the B2 level requires HLC competence (Hulstijn, 2011). Hulstijn argues that there will be substantial differences in how L1 users perform HLC tasks and his second corollary predicts a relatively wide range of scores when L1 users complete such tasks.

Existing research into L1 performance on L2 writing tasks supports these hypotheses. Two such studies were undertaken in Flanders. In the first, two L2 writing tasks were administered to 176 first-year students of four Flemish university colleges (Van Houtven & Peters, 2010). The results indicated that not all respondents reached the threshold level, but since they differed greatly in terms of educational background the authors hesitated to draw any firm conclusions. A more recent paper (De Wachter et al., 2013) confirmed that there is large variation in the Dutch language proficiency of first-year students at a Flemish university.

A few studies have considered native speaker test performance on English language tests for university admission. Hamilton et al. (1993) analyzed the test scores of native speakers (N = 48, 32 bilingual, 16 monolingual) on IELTS writing tasks (one information transfer task and one argumentation task), and found substantial variation in the L1 performances. Noting that the mean writing score from L1 test takers was around IELTS 6.5 (i.e., B2), the authors concluded that not all L1 test takers met the threshold demanded of L2 students, and questioned what this meant for equity of access to university. Focusing on the TOEFL iBT, Stricker (2004) used t-tests and Cohen's d to establish that 168 "American-born speakers of English" (no information on home language use was offered) significantly outperformed L2 test takers on an essay writing task ($p < .05$). Furthermore, he found that even though the score variance in the American population was significantly smaller ($p < .05$) than the L2 scorer variance, it was not unsubstantial, and not all L1 users passed the L2 threshold.

The above-mentioned studies suggest that not all students who are considered native speakers meet the linguistic requirements demanded from L2 students upon university entrance. The next segment of the literature review

focuses not so much on the question if L1 and L2 students' writing skills differ, but on how they differ.

Writing performance differences

Comparing L1 and L2 texts using a four-dimension rating scale, Weigle and Friginal (2015) only found one significant difference on the four dimensions they had identified, namely expression of opinion: L2 learners used significantly more opinion statements than L1 users. Weigle (2002) reported that scores for vocabulary are often found to be the strongest predictor of L1 background, and other studies have also found vocabulary to be the strongest predictor of overall scores in L1 (e.g. Wolfe, Song, & Jiao, 2016) and L2 (Koda, 1993) writing tasks. Findings for grammatical accuracy show that L1 users usually outperform L2 learners in this respect (Leki, Cumming, & Silva, 2008), while findings pertaining to syntactic complexity are less clear-cut. Huie & Yahya (2003) reported that L1 users write more complex sentences, while Lee (2003) found no such differences. In this respect, Schoonen et al. (2003) suggested that L1 writing may depend more on metalinguistic and topical knowledge, while L2 writing proficiency could be better explained in terms of linguistic knowledge.

L1/L2 writing research has also explored how L1 and L2 writers engage with writing tasks. It is clear that L1 and L2 writing processes are different, yet not entirely disconnected (Polio, 2013; Schoonen et al., 2003). Skilled L2 writers tend to be skilled L1 writers too, if they have surpassed a certain threshold proficiency level (Leki et al., 2008). Still, perhaps because writing in a language that is not one's L1 requires mental capacity (Leki et al., 2008), L2 writers appear to be less flexible in how they reply to writing prompts. L2 writers tend to adopt a more systematic approach (Van Weijen, Van den Bergh, Rijlaarsdam, & Sanders, 2008; Victori, 1999), and appear to adhere more rigorously to the task instructions. Because they tend to stick closely to a fixed routine or scenario, L2 writers have also been found to adopt a smaller range of writing strategies across different writing tasks (Rijlaarsdam et al., 2005).

Research has not only focused on differences between L1 and L2 users, but also on different types of L2 learners. Some studies have compared L2 writers by proficiency level, while others compared L2 learners who acquired the language at home with those who attended a study abroad program. Research focusing on the first dichotomy has found that skilled L2 writers tend to be more fluent writers (de Larios, Marín, & Murphy, 2001), better at planning (Victori, 1999), and more focused on text level, whereas less skilled L2 writers tend to focus more on vocabulary and grammar (Leki et al., 2008; Victori, 1999).

Next, the study abroad (SA) literature offers further insights that are relevant in the light of the current study. Importantly, the term "Study Abroad" can cover a whole range of experiences (Engle & Engle, 2003) from two-week

immersion programs, to students pursuing a degree abroad, where the L2 is the medium of instruction. Llanes et al. (2012), focusing on Erasmus students in Europe, found that learning a language in a SA context does not automatically lead to better writing skills, and that some SA experiences actually lead to less L2 learning than in-class instruction at home (Llanes et al., 2012). Also, studying the effects of an Erasmus experience on L2 development, Serrano et al. (2012) found that the positive effects of SA extend less to the written modality than to the oral and the lexical domain. One consistent finding in the SA literature is that L2 learners who study a language in a target context do not always have many meaningful interactions with native speakers. This was observed in a number of studies, such as Amuzie and Winke, (2009) – who tracked the experiences of Chinese and Korean students enrolled at North American universities – and Gu and Maley (2008) who focused on the experience of Chinese students at British universities.

As far as could be determined, no study abroad research has yet compared the performance of SA (study abroad) and AH (at home) L2 learners on centralized high-stakes tests. In the broader field of ESL, however, a limited number of studies have done so. Gu (2014) found no difference in TOEFL iBT scores for L2 learners who learnt English in an English-speaking context or in an at-home context – a finding that contradicts an older study by Ginther & Stevens (1998).

Another useful categorization within the broader group of L2 learners distinguishes international L2 learners from G1.5 learners (Harklau et al., 1999). The former group is enrolled in post-secondary education after learning the L2 in a classroom in their home country, while the latter attended secondary education in the L2 context before moving on to post-secondary education (di Gennaro, 2013). There is relative agreement in the literature (di Gennaro, 2016 offers an overview) that both groups score differently on grammatical criteria, but the findings regarding vocabulary and other criteria are less uniform. A recent study (di Gennaro, 2016) found no significant differences in scores between international L2 and G1.5 leaners on the scores for cohesion, rhetorics, sociopragmatics, and content. Differences in grammar score – the hardest criterion for both groups – did reach significance, however. The current study does not directly focus on G1.5 learners, but will refer to the research literature when relevant.

RESEARCH QUESTIONS

The literature review above strongly suggests that it is unlikely that all L1 students have obtained the language proficiency level that is required of international L2 students at the start of their academic studies at a university. No study has yet compared L1 and L2 performances on a centralized high-stakes university entrance test with regard to B2 proficiency. Finally, little is known about the performance of different types of L2 users on a high-stakes test. Therefore, this study explores whether all L1 students attain a B2 level on an L2 entrance test (Assumption 4) and goes on to compare L1 students' and L2 learners' performance on the same L2 entrance test, and to investigate the performance of SA ($L2_F$) and AH ($L2_I$) learners.

By comparing the performance of three groups of first-year university students, this large-scale between-group study addresses a number of gaps in L2 writing assessment research. The study is relevant because it could corroborate Hulstijn's second corollary ("Individual differences among adult L1ers will be relatively large in tasks involving HLC discourse, in all four modes of language use" – Hulstijn, 2015, p. 25). Furthermore, this study adds a new dimension to the study abroad literature, which to date has mainly focused on qualitative differences between SA and AH students rather than on their scores on centralized tests. Finally, by including L1 performance on L2 tasks, this study may have implications for the university entrance policy in Flanders. The implicit hypothesis underpinning this policy is that a B2 level is a prerequisite for successful participation in academic life, and that Flemish students possess this level after graduating from secondary school.

This study aims to answer two primary research questions. RQ1 is concerned with Assumption 4:

RQ1 *Do all students with a Flemish high school degree, who are exempt from taking a university entrance language test, pass the B2 threshold*?

The hypothesis was that not all students who graduated from a Flemish secondary school would pass the B2 threshold as measured by STRT. This hypothesis is in line with Hulstijn's (2015) BLC/HLC theory and with previous studies on L1 performance (De Wachter et al., 2013; Hamilton et al., 1993; Stricker, 2004; Van Houtven & Peters, 2010).

RQ2 *What are the differences between the writing performances of Flemish high school graduates and international L2 students?*

2a. *Do Flemish students outperform international L2 candidates on the STRT writing tasks?*

Based on results from previous research (Weigle, 2002; Weigle & Friginal, 2015) it was assumed that Flemish students would outperform L2 students, but not necessarily on every criterion (Huie & Yahya, 2003; Lee, 2003; Leki et al., 2008).

2b *Are there any performance differences between L2 learners who learned Dutch in Flanders and those who did so at a language school outside of Flanders?*

Because of the inconclusive results in the study abroad literature, no predictions were made regarding the performance of either group.

PARTICIPANTS & METHODOLOGY

STRT Writing tasks

Since ITNA does not include any writing tasks at the B2 level, it could not be used to compare writing performances, so it was not considered as a measurement instrument in this study. STRT consists of six integrated tasks (see Appendix 1), which are intended to reproduce real-life situations and activities in the context of higher education in Flanders, where writing becomes progressively more important during a students' academic career (De Wachter et al., 2013; Herelixka, 2013).

Task selection

Because of time constraints it was not feasible to administer all test tasks to the L1 population. Consequently, the two STRT writing tasks that accounted for most of the variance in the writing scores (Table 6.1) were selected. Based on all available STRT scores (N = 913) a regression model was constructed to identify the writing tasks that explained most score variance on the written component. In this model, Task 2 (T2) and Task 4 (T4) turned out to be the strongest predictors of the overall STRT score. A regression model composed solely of Task 2 and Task 4 explained 91% of the score variance in the written component of STRT (R^2_{Adj} = .911 (F (115) = 599.9, p < .000) T2 β = .602, T4 β = .451).

Table 6.1. Multivariate linear regression: STRT written score ~ T1-T4 scores

	B(SE)	*β*
(Intercept)	-0.015 (.01)	
T1	2.377 (.001) ***	.245
T2	2.475 (.001) ***	.392
T3	2.772 (.001) ***	.254
T4	2.380 (.001) ***	.328

Note. R^2 Adjusted = 1, $p < .000$

T2, a listening-into-writing task, requires test takers to listen to a scripted nine-minute lecture about industrialization. Candidates listen to the lecture twice while taking notes. Afterwards they have thirty minutes to write a summary. In T4, a reading-into-writing task, test takers have one hour to summarize a text about the pros and cons of schools without gender differentiation in the teacher corps, and to formulate their own opinion about the topic.

Integrated writing-from-reading or writing-from-listening tasks of this kind have been said to tap into an essential part of academic language use (Chan, Inoue, & Taylor, 2015; Cumming, 2013). Writing-from-reading is considered an important index of academic achievement (Baba, 2009; Hirvela, 2016), and summary tasks in particular offer valuable information in this respect. Similarly, note-taking in writing-from-listening tasks represents an important part of what students need to be able to do (Lynch, 2011; Song, 2012). STRT developers consider writing tasks a good yardstick for determining whether a candidate will be able to meet the linguistic demands of university (Confidential STRT document, 2014). Appendix 1 includes a more thorough description of T2 and T4, and describes the rating criteria used to assess writing performances.

Scoring

All performances were independently double rated by 17 trained STRT raters who were unaware of the background of the candidates, in order to avoid L1 bias (Van Weijen, 2009; Weigle, 2002). The rater reliability in this study was satisfactory (exact vs. expected agreement 63.9% - 56.8%; Infit MnSq μ = .99; X^2 (17) = 17.9, p = .40). The introduction and Chapters 3 and 4 contain more information about STRT rater training and selection.

Participants & data collection

Three groups of participants were involved in this study. Each group took the test tasks under examination conditions. Trained examiners and invigilators were

always on site to ensure that the examination conditions complied with instructions in the test manual.

The first group consisted of 159 first-year students of business studies who had graduated from Flemish secondary education. In return for their collaboration, these students were awarded credits for a curricular course at the start of the year. In the light of the first research question it was important that the demographic composition of this group represented that of a typical first-year course, minus the international students. Since the first research question is whether all students who are exempt from taking a binding language test would pass its threshold, the main selection criterion for this population was having a Flemish secondary school degree. Eleven percent of the Flemish participants had a home language different from Dutch, but for all members of this group, Dutch was their language of schooling. The Flemish population thus consists of two subpopulations, L1 (i.e., home language and language of schooling is Dutch) and G1.5 students (i.e., the students' home language is different from Dutch, but they graduated from a Dutch-medium secondary school in Flanders). Since the size of the G1.5 population was rather limited (n = 18), and since this group was not part of the research questions guiding this study, performance data of the G1.5 population will not be a primary focus of this chapter, but references to the relevant literature will be made.

The second column of Table 6.2 ("Flemish") displays the demographic variables of the Flemish research population, 85% of whom had attended the academic strand of secondary education (the other 15% had attended the technical strand of secondary education). The demographics are representative of first-year university students in Flanders (58% female; median age 18; 10% migration background; 96% no previous university experience; 90% academic strand of secondary education. See Glorieux, Laurijssen, & Sobczyk, 2015; Universiteit Gent, 2013).

Table 6.2. Demographic data of participants

		Flemish	$L2_F$	$L2_I$
Age	Average (SD)	18 (.7)	27 (7)	20 (5)
	Min -Max	17-21	16-50	14-55
Gender: Female		50%	70%	63%
L1		89% L1 Dutch	34% Italic	57% Italic
		3% Italic	20% Balto-Slavic	18% Germanic
		3% Balto-Slavic	11% Germanic	6% Balto-Slavic
		5% Other	35% Other	19% Other
University experience		10%	59%	≤32%

The Flemish participants performed the two written STRT tasks at the beginning of the second month of the academic year 2015-2016. L1 respondents received no specific training or preparation, as the skills required for successful task performance are in line with the attainment targets of the academic and technical strands of secondary education issued by the Flemish government (Onderwijs Vlaanderen, 2015). Prior to taking the test, Flemish respondents filled out a form with basic demographic information (gender, age, L1, educational background).

$L2_F$ users are one of two groups of international L2 students considered in this chapter. Throughout the dissertation, the code $L2_F$ (i.e., L2 learners in Flanders) is used to refer to test takers who learned Dutch in Flanders and took the test there. In this chapter, the $L2_F$ population is somewhat larger, because scores collected in the regular Flemish May 2015 STRT administration were included, in addition to the test scores of $L2_F$ candidates (N = 118) who had taken both STRT and ITNA used in Chapters 3 and 4. The total number of $L2_F$ participants included in the analyses of this study, is 168. These L2 learners had not attended any Dutch classes prior to arrival in Flanders. No members of this population had attended secondary school in Flanders. The median length of Dutch L2 instruction for this group of respondents was eighteen months. Information regarding prior university experience was available for 118 respondents. The respondents in this group received no specific in-class preparation for STRT, but they were familiar with writing tasks and they had received a link to sample tasks on the STRT website. This group of respondents filled out the same demographic information form as the Flemish students.

The third group of participants in this study consists of the regular STRT population: L2 learners who had studied Dutch at a language school in their home country, and took the entrance test there. These candidates received the standard STRT demographic information sheet (gender, age, L1), but due to matters of privacy and data protection, it was impossible to trace which of these candidates had enrolled where. Even though the $L2_I$ data do not include direct information concerning prior education, we can assume that at least 68% had no university experience at the time of data collection, because they were eighteen or younger when they took the test. Candidates belonging to this dataset (N = 526) will be referred to as $L2_I$ (i.e., L2 learners who studied Dutch at an international language school outside of Flanders).

Since the $L2_I$ population represents the full test-taking population of the May 2015 STRT administration, this dataset is substantially larger than the other two. For methodological reasons, it was decided not to balance the sample sizes however. Undersampling – reducing the size of the largest set by randomly removing cases – has major drawbacks, the primary one being the loss of potentially valuable data (Kotsiantias et al, 2005). Oversampling on the other

hand – randomly duplicating cases from the smaller dataset – may entail analogous risks, such as overfitting data (Kotsiantias et al, 2005). Consequently, instead of randomly adding or removing observations to the datasets, only those analyses that allow for unbalanced datasets were conducted.

Data analysis

RQ1 *Do all students with a Flemish high school degree, who are exempt from taking a university entrance language test, pass the B2 threshold?*

Descriptive statistics of the combined scores of T2 and T4, and a Multi-Faceted Rasch analysis (MFRA) were used to determine whether all Flemish students passed the STRT writing tasks. A binomial sign test was performed to test the null hypothesis (P_{L1}^{Pass} = 1). For the MFRA four variables were identified: candidate ability, rater severity and item difficulty. The variable group (L1, $L2_I$, $L2_F$) was entered as a dummy facet. T1 and T3 scores were treated as missing data for the L1 group. L2 scores on the oral component were available but were excluded from the analysis because no spoken L1 performances were available and oral and written performances on LAP tests have been shown to belong to different dimensions (Gu, 2014). In line with the STRT procedure, this study adopted a Rasch measure of 1.42 as a cut-off point for pass/fail decisions. This cut score was determined in a standard setting procedure (CNaVT, 2014). Also following STRT procedure, a pass judgment was assigned to candidates whose score was within the range of one standard error below the cut score.

RQ2a *Do Flemish students outperform international L2 candidates on the selected STRT writing tasks?*

A Wilcoxon signed-rank test was performed to determine whether any differences in median scores on content and form criteria between the different populations were significant. The same analysis was conducted within the Flemish sample, to compare the scores of L1 and G1.5. Given the small proportion of G1.5 students (11%), no further analyses were conducted within the Flemish group.

A principal component analysis (PCA) was conducted on the standardized z-scores to determine which criteria could be combined for the Wilcoxon signed-rank test. The PCA was run on the ten rating criteria using oblique promax rotation, since the variables were known to be correlated. Bartlett's test of sphericity showed that a PCA was warranted (X^2 (45) = 3878, p < .000), and the Kaiser-Meyer-Olkin measure showed the sampling adequacy (KMO = .87) to be good (Kaiser, 1974). All individual criteria had KMO values (> .83) above the .5 limit (Field, Miles, & Field, 2012). Based on the initial analysis and the scree plot,

it was decided to run the analysis with three factors, as three components had eigenvalues at or above one, which – taken together – explained 70% of the score variance. Table 6.3 shows the factor loadings of the different criteria and indicates that it was warranted to combine the linguistic criteria of T2 and those of T4. The content criteria of both tasks load onto the third component.

Table 6.3. Promax rotated factor loadings

	Form T2	Form T4	Content
T2 *Vocabulary*	.86		
T2 *Grammar*	.83		
T2 *Cohesion*	.77		
T2 *Spelling*	.67		
T4 *Vocabulary*		.9	
T4 *Grammar*		.78	
T4 *Cohesion*		.69	
T4 *Spelling*		.65	
T2 *Content*			.83
T4 *Content*			.81
Eigenvalues	2.74	2.65	1.59
% of variance	27	27	16

Note. Factor loadings <.3 were omitted from the table.

In order to better understand how L1 candidates performed on the separate criteria, a logistic regression model was developed in which L1ness was operationalized as a function of the rating criteria. All criteria were individually included in the regression, because such a model was a significantly better predictor than one which used the clustered criteria (X^2 (4) = 159, p < .000).

RQ2b *Are there any performance differences between L2 learners who learned Dutch in Flanders and those who did so at a language school outside of Flanders?*

Multinomial logistic regression was used to identify how well rating criteria predicted membership of one of the three respondent groups: L1, $L2_I$ (= AH) or $L2_F$ (= SA). Logistic and multinomial regression models were also used to measure the effect of other background variables (age, gender, L1, educational background, cf. Weigle, 2002) on scores. Since these results indicate the extent to which performance differences between research populations have been influenced by background variables, but do not directly answer the research questions, they are briefly reported here.

The effect of educational background was examined using the $L2_F$ population only, because students with no experience in higher education were

overrepresented in the L1 group, and no educational background information was available for the $L2_I$ students (for the other bias analyses, the full population data were used). Having university experience was a significant positive predictor for L2F scores on *Cohesion* and *Vocabulary*, but a negative predictor for *Grammar* (*Cohesion* T2 (*B(SE)* = 1.23(.43)**; *Vocabulary* T4 (*B(SE)* = 1.05(.43)*; *Grammar* T4 (*B(SE)* = -1.04(.45)*). Given the sample size available for this background variable, generalizations (Nagelkerke Pseudo R^2 = .18) should be made cautiously.

Since the dataset included nineteen different language branches, languages were clustered into five groups: "Germanic", "Italic", "Balto-Slavic", "Other Indo-European" and "Other". An exploratory logistic regression showed that overall scores predicted membership of the Germanic language group (*B(SE)* = 0.26(.06)**), so "Germanic" was used as the baseline in the multinomial logistic regression. The model with two additional predictors ("Italic" and "Other") yielded the most explained variance (McFadden R^2 = .18, X^2 = 290, p < .000) and is reported here. MFRA reliably (.92) determined that Germanic/Italic candidates had higher test scores than other test takers. Overall, scores on linguistic criteria did not significantly predict membership of the Germanic/Italic group, but scores on content criteria did (T2 *B(SE)* = 0.26(.06)***; T4 *B(SE)* = 0.23(.06)***).

Additionally, Facets bias analyses on task level were conducted for each background variable. In each MFRA model, the background variable was entered as dummy facet. The regression models and the MRFA both showed that candidates aged eighteen and younger significantly outperformed older ones (*B(SE)* = -.28(.032)***; MFRA reliability .96). Other bias analyses yielded non-significant or minute differences.

RESULTS

Do all L1 students pass the B2 threshold, as measured by STRT?

The binomial sign test rejected the null hypothesis at the 95% and 99% confidence level (p < .000). Not all Flemish candidates pass the STRT writing tasks. As a group, Flemish students perform best, but not all make the grade.

Table 6.4 shows that the median and mean scores of the Flemish group were higher than those of both L2 groups. This does not imply that individual Flemish test takers performed best on the test, however, since the best Flemish performer was outperformed by 21 $L2_I$ and $L2_F$ candidates. Overall, however, the Flemish group gained the highest scores, and the range between the highest and the lowest score is the smallest in the Flemish group. On a twenty-point scale, the range of scores on T2 and T4 combined is 7.8 for the Flemish group, versus 15.5 and 18 for the L2I and the L2F group respectively.

Table 6.4. Descriptive statistics: group scores, combined scores T2 and T4

	Flemish	$L2_F$	$L2_I$
N	159	168	526
Mean	15.74	12.49	14.52
SD	1.62	3.58	2.09
Median	15.71	13.06	14.69
Min	10.61	1.22	4.08
Max	18.37	19.18	19.59
Range	7.76	17.96	15.51
Skew	-0.56	-0.8	-0.45
Kurtosis	0.04	0.47	0.97
SE	0.13	0.28	0.09

Note. Max score = 20

The MFRA analysis for the facet "Group" (see Table 6.5) confirmed the trends emerging from the descriptive statistics: in general, L1 candidates were the strongest candidates on the STRT writing tasks. The Rasch model reliably (.99) separated Flemish, $L2_I$ and $L2_F$ candidates in terms of ability, implying that each group performed at a distinctly different level. The L1 group outscored both L2 groups, but not all L1 candidates scored above the STRT cut off point.

Table 6.5. MFRA for facet "Group"

	Measure	Model SE	Infit	% below cut-off
Flemish	0.57	0.02	1.12	11
$L2_I$	-0.15	0.01	0.82	30
$L2_F$	-0.43	0.01	1.29	57

Eleven percent of the Flemish group scored below the required 1.42 threshold and did not pass the writing tests. In comparison, 30% of the $L2_I$ population and 57% of the $L2_F$ population did not obtain the required score. The ability measure (reliability .99) of the Flemish population spanned four logits, compared to five for $L2_F$ and six for $L2_I$ candidates, reaffirming that the range of scores for the Flemish participants was smaller than for the other groups.

The demographics of the small group of Flemish candidates who scored below the cut score showed a slight overrepresentation of men, of G1.5 students, and of students who attended a technical strand of secondary education. One out of five G1.5 students failed the writing test, versus one out of ten Flemish L1 students. One in three students with a technical education background failed the writing test. Six out of ten Flemish students who failed the test were male.

What are the differences between the writing performances of Flemish high school graduates and international L2 students?

Wilcoxon signed-rank test results (see Table 6.6) showed that Flemish students consistently outperformed both L2 groups on the linguistic criteria (max score = 16), but the effect size *r* was the largest for L1 versus $L2_F$ candidates. The content scores (max score = 17) revealed that Flemish students performed similarly to $L2_F$ candidates, and both groups' median scores were below the $L2_I$ median, with medium effect sizes.

Table 6.6. Wilcoxon signed-rank test: Flemish, $L2_F$ and $L2_I$

	Median			Flemish & $L2_F$			Flemish & $L2_I$			$L2_I$ & $L2_F$		
	Fl	$L2_F$	$L2_I$	*W*	*p*	*r*	*W*	*p*	*r*	*W*	*p*	*r*
T2Ling	13	9.5	10	23275	<.000	-.64	71833	<.000	-.53	53196	<.000	-.15
T4Ling	13.5	10	11	22750	<.000	-.61	67486	<.000	-.45	53684	<.000	-.16
Cont	12.5	12	15	14825	.08	-.09	23028	<.000	-.33	62871	<.000	-.32

Wilcoxon signed-rank tests were also carried out within the Flemish sample to determine whether there were any performance differences between Flemish L1 candidates and G1.5 students (see Table 6.7). The effect sizes show that the differences are rather small, or – in one case – non-significant. The differences between the $L2_F$ and the G1.5 scores on linguistic criteria, however, are large and significant, and a detailed look at the individual criteria shows that G1.5 candidates significantly ($p < .05$) outperformed the $L2_F$ candidates on every linguistic criterion. The effect sizes were the largest for *Vocabulary* (T2 *r* = -.30; T4 *r* = -.30) and *Grammar* (T2 *r* = -.36; T4 *r* = -.24).

Table 6.7. Wilcoxon signed-rank test: L1, G1.5, and $L2_F$

	Median			L1 & G1.5			$L2_F$ & G1.5		
	L1	G1.5	$L2_F$	*W*	*p*	*r*	*W*	*p*	*r*
T2Ling	13.5	12.5	9.5	827.5	<.05	-.18	2562.5	<.000	-.35
T4Ling	13	12.5	10	916.5	.07	-.14	2369	<.000	-0.29
Cont	12.5	11.25	12	794	<.05	-.19	1344	.43	-.05

Note. r = effect size

In order to determine whether Flemish students outperformed L2 students on all linguistic criteria, a logistic regression model (see Table 6.8) was constructed to predict membership to the Flemish group from criterion scores (Nagelkerke Pseudo R^2 = .41). In this model, *Grammar* and *Vocabulary* were the only significant positive predictors of Flemish group membership, whereas content scores were significant negative predictors. The *Content* score of T2 was the

strongest (negative) predictor of this model. *Cohesion* and *Spelling* did not significantly predict whether or not candidates belonged to the Flemish group.

Table 6.8. Logistic regression: Flemish ~ criterion scores

	B(SE)		*z*	*Odds ratio*
(Intercept)	-7.33	(0.79)***	-9.3	
T2 *Vocabulary*	1.018	(0.22)***	4.62	2.768
T2 *Grammar*	0.503	(0.22)*	2.29	1.654
T2 *Cohesion*	0.051	(0.18)	0.29	1.053#
T2 *Spelling*	-0.013	(0.2)	-0.06	0.987#
T2 *Content*	-0.382	(0.07)***	-5.45	0.682
T4 *Vocabulary*	1.151	(0.23)***	4.95	3.162
T4 *Grammar*	0.571	(0.23)*	2.43	1.769
T4 *Cohesion*	-0.271	(0.2)	-1.33	0.762#
T4 *Spelling*	0.132	(0.18)	0.72	1.141#
T4 *Content*	-0.102	(0.07)	-1.41	0.903#

Note. * $p < .05$; ** $p < .01$; ** $p < .001$

confidence interval crosses 1

In order to compare both L2 groups with each other, while allowing for comparison with the Flemish candidates, the multinomial logistic regression model (McFadden R^2 = .37, Likelihood ratio test X^2 = 587.67, p < .000) was run twice; once with the $L2_I$ group as a baseline, and once with the Flemish population. Table 6.9 displays the results of $L2_F$ vs $L2_I$, $L2_I$ vs Flemish, and $L2_F$ vs Flemish.

The first part of Table 6.9 (i.e., $L2_F$ vs $L2_I$) indicates that the only criteria that significantly predict whether a candidate is $L2_I$ or $L2_F$ were *Vocabulary* in T2, cohesion in T2, and content in T2 and T4. *Vocabulary* scores in T2 can be considered a positive predictor of $L2_F$ membership, while *Cohesion* scores on Task 2, and *Content* scores on both T2 and T4 tasks are negative predictors. A one-unit increase in *Vocabulary* scores can be expected to increase the odds of belonging to $L2_F$ over $L2_I$ by 0.92 units. A one-unit increase in the *Cohesion* and *Content* scores will decrease those odds.

When considering the second and third section of Table 6.9, two similar profiles emerge: *Vocabulary* is a negative predictor of both L2 groups, but the opposite is true of *Content* scores. Higher content scores increase the odds of belonging to either L2 group, but the trend is most pronounced for $L2_I$ candidates. Higher *Vocabulary* scores on T2 and T4 are associated with Flemish candidates, but for the other linguistic criteria, the results are more heterogeneous. *Grammar* is a positive predictor of membership to the Flemish group only in T2, and *Spelling* only in T4. Remarkably, *Cohesion* is a positive L2-predictor in T4, but a significantly negative one for $L2_F$ candidates in T2.

Grammar in T4 and spelling in T2 did not significantly predict membership of any of the three respondent groups.

Table 6.9. Multinomial logistic regression

	$L2_F$ vs. $L2_I$		$L2_I$ vs. Flemish		$L2_F$ vs. Flemish	
	B(SE)	*Odds ratio*	*B(SE)*	*Odds ratio*	*B(SE)*	*Odds ratio*
(intercept)	2.68 (0.55)***		9.86 (1.29)***		12.54 (1.32)***	
T2Vocabulary	0.92 (0.21)***	2.50	-2.26 (0.33)***	3.84	-1.34 (0.35)***	0.40
T2Grammar	-0.30 (0.20)	0.74#	-0.62 (0.31)*	2.51	-0.92 (0.33)**	1.35
T2Cohesion	-0.45 (0.19)*	0.64	-0.26 (0.25)	2.03#	-0.71 (0.28)*	1.57
T2Spelling	0.00 (0.17)	1.00#	-0.04 (0.30)	1.04#	-0.04 (0.32)	1.00#
T2Content	-0.41 (0.07)***	0.67	0.84 (0.10)***	0.65	0.43 (0.10)***	1.50
T4Vocabulary	0.34 (0.20)	1.40#	-2.35 (0.35)***	7.46	-2.01 (0.37)***	0.71
T4Grammar	-0.28 (0.20)	0.76#	-0.20 (0.32)	1.62#	-0.48 (0.33)	1.32#
T4Cohesion	-0.12 (0.19)	0.89#	1.01 (0.30)***	0.41	0.89 (0.32)**	1.13
T4Spelling	0.01 (0.16)	1.01#	-1.02 (0.28)***	2.75	-1.01 (0.30)***	0.99
T4Content	-0.25 (0.07)***	0.78	0.55 (0.11)***	0.74	0.30 (0.11)**	1.28

Note. * $p < .05$; ** $p < .01$; *** $p < .001$
confidence interval crosses 1

DISCUSSION

The data presented in this study show that eleven percent of the Flemish participants do not meet the written language demands that their international L2 peers are required to meet. The Flemish group as a whole was the most successful, but the best performers on the writing tasks were L2 students. No Flemish candidates appeared in the highest-scoring percentile. Even though the score range was wider in the L2 groups, there was substantial variation in the

scores of the Flemish candidates (a range of 7.8 on a twenty-point scale). The group of L1 students who did not pass the writing tasks is too small to draw any firm conclusions, but the trends in the group composition reflect the results of large-scale research into performance indicators in Flemish post-secondary education (Glorieux et al., 2015). In absolute numbers, most L1 candidates who did not attain the B2 level, were monolingual Dutch students (n = 13) who had graduated from the academic strand of secondary education (n = 11), but proportionally, students with a home language other than Dutch, and students with a degree from the technical strand of secondary education are slightly overrepresented (Glorieux et al., 2015). The overrepresentation of G1.5 students among low-scoring Flemish test takers echoes Fox (2005), who concluded that it would be wrong to assume that G1.5 students will perform equally well in the target context as their L1 peers.

This study also supports Hulstijn's claim that some L2 learners will outperform L1 users on cognitively demanding tasks, and it is in line with previous findings by Hamilton et al. (1993) and Stricker (2004). The data also back Hulstijn's hypothesis that there will be relatively large differences in the performances of L1 users on cognitive tasks (Hulstijn, 2015, p. 53), even though the Flemish population in this study was relatively homogenous in terms of age, home language and secondary school degree.

The fact that more than one out of ten Flemish participants failed to meet the writing demands of a B2 university entrance test could have considerable implications. The results indicate that the differential treatment of Flemish students and L2 students (with regard to university entrance tests) may be based on an assumption that is not supported by empirical data or by recent theories: not all students who graduated from a Flemish secondary school possess the B2 level, as measured by STRT. Additionally, previous research has shown that allowing G1.5 students to enroll without taking a language test, may not always be in their best interest, since these students may be less likely to achieve academic success than their L1 peers (Fox, 2005).

Even though the best $L2_I$ and $L2_F$ candidates outperformed the best Flemish test takers, the analyses showed that the Flemish group as a whole did better than the L2 test takers in terms of overall scores and scores on linguistic criteria. The linguistic criteria that significantly predicted membership of the Flemish subpopulation were vocabulary and, to a lesser extent, grammar and spelling. The scores on cohesion did not offer a clear picture: on one task, high cohesion scores were a positive predictor of L2 groups in contrast to the Flemish group, while on the other task high cohesion scores significantly predicted membership of the Flemish population rather than the $L2_F$ group.

A comparison of the scores of the Flemish and the L2 populations shows two main trends: Flemish students, both L1 and G1.5, scored significantly higher on linguistic criteria, and L2 students gained higher content scores. The scores

for content are perhaps the most surprising, all the more because within the Flemish sample, L1 students significantly outperformed G1.5 students on content, but not on the linguistic criteria of both tasks.

If L1 writing performance depends primarily on content and metacognition (Schoonen et al., 2003) and if vocabulary is a key component in listening and reading, one would expect L1 users to have an advantage over L2 learners when it comes to scoring content points. One explanation may be found in the academic background of the L2 population, but bias analysis showed that this variable did not have a significant impact on content scores. L2 learners' test-taking strategy could provide an alternative explanation for content scores. L2 writers prefer to stay closer to the source material than L1 writers (Keck, 2006, 2014; Wu, 2013) and tend to stick closely to a fixed routine or scenario (Rijlaarsdam et al., 2005). The binary content criteria in STRT actually award candidates who stick closely to the prompt and directly – though not literally – reuse elements from the input material. This could explain why the L2 group performed better than the Flemish population on STRT content criteria. The results of the content criteria could be a cause of concern for STRT developers, and urge them to examine the construct relevance of the operationalization of content items.

The Flemish group outperformed both L2 groups on linguistic criteria. The PCA showed that linguistic criteria cluster together on a task basis, rather than by criterion. This reflects the data found in Tillema et al. (2013) and may indirectly confirm Yu's (2013a, 2013b) hypothesis that input material used in integrative tasks may impact performance more than a candidate's writing ability. The PCA also confirms Schoonen et al. (2003) who argue that the score on a writing task should be seen as the interaction between various linguistic features such as grammar, vocabulary and structure. Nevertheless, the data reaffirm that *Vocabulary* scores are an important predictor of overall writing scores (Koda, 1993; Weigle, 2002; Wolfe et al., 2016), and Flemish candidates performed significantly better on this criterion. Reminiscent of previous research (di Gennaro, 2016), the G1.5 subpopulation outperformed the L2F population on every linguistic criterion, but the largest effect sizes were measured for *Vocabulary* and *Grammar*. In di Gennaro's study too, G1.5 participants outperformed international L2 students in terms of word choice and in terms of grammatical criteria.

Overall, $L2_I$ candidates outperformed their $L2_F$ peers. Studying Dutch in Flanders (i.e., abroad) does not seem to automatically lead to higher scores on the written component of STRT than studying Dutch in the home country. The vocabulary score on the writing-from-listening task (T2) was the only positive predictor for $L2_F$ candidates in the multinomial regression model, which is in line with Leki et al. (2008), who observed that less skilled L2 writers tend to focus on

lexis. *Content* scores on both tasks significantly predicted membership of the $L2_I$ group. The other criteria did not significantly contribute to the model.

At first sight it may seem striking that learning Dutch in a context that offers great opportunities to become immersed in the target language environment may not pay off in terms of test outcomes, but similar results have been found in previous studies. Research has shown that a study abroad experience does not necessarily benefit overall writing proficiency, while it may lead to gains in oral fluency or lexical development (for an overview see Sanz, 2014). Reminiscent of previous studies, the results for *Vocabulary* reported here showed that $L2_F$ candidates gained higher vocabulary scores than L2I students (Housen et al., 2011; Juan-Garau, Salazar-Noguera, & Prieto-Arranz, 2014). This seems to confirm that study abroad is more beneficial to lexical development than learning a language in one's native country. Since this study did not focus on measuring progress or gains, further presumptions or generalizations on this matter would be speculative at best.

$L2_I$ students did better than their $L2_F$ peers on content criteria. This group may have stuck more closely to the source text (see above) and may have been awarded for it in the rating process. Alternatively, even though L2 users as a group tend to stick closer to source texts than L1 users, L2 candidates with a higher proficiency incorporate more details from the source text in their performance than their less proficient peers (Wu, 2013). Possibly, the $L2_I$ group was more proficient and more able to include more content specificity.

A further explanation for the difference in scores may lie in the length of instruction. The L2I candidates in this study had typically been learning Dutch for a few years, while the median length of Dutch instruction for $L2_F$ candidates was eighteen months. Possibly, candidates who have been studying Dutch longer tend to score better on STRT.

CONCLUSION: ASSUMPTION 4

This chapter has found no supporting evidence for Assumption 4. Even though as a group Flemish students outperformed both groups of L2 students, 11% did not meet the B2 level requirements as measured by STRT. Based on these data, it cannot be automatically assumed that Flemish high school graduates who are exempt from university entrance language tests will perform up to criterion.

PART 3
GAINS & CONTEXT

> The first two parts of this dissertation were concerned with empirically investigating assumptions that support the university entrance policy for international L2 students. Until now, the focus was on proficiency levels, on representativeness, and on test equivalence. In the third part, which consists of one chapter, the attention shifts to what happens after the entrance test: do L2 students make gains in terms of academic Dutch language development, and if yes or no, why?

The research goal examined in this part (RG 5) does not rely on an implicit or explicit assumption present in policy texts. Instead, it examines whether the assumptions relating to language gains voiced by some individual policy makers (see Chapter 2) hold true. Additionally, since the first part of the dissertation showed that most international L2 students are not entirely ready for the linguistic requirements of academic studies at university, and that the B2 level offered no guarantees for coping linguistically in the TLU context, post-entry language gains would benefit international L2 students. To date however, there was no information available to show whether or why these gains are made.

Chapter 6 is based on:

> Deygers, B. (2017, accepted). A year of highs and lows. Considering contextual factors to explain L2 gains at university. *The Modern Language Journal.*

This paper has been edited to avoid redundancy. The method section has been revised.

CHAPTER 6
EXAMINING L2 GAINS

> The exponential increase of international L2 students at universities worldwide has generated substantial research interest into university entrance language requirements and university entrance language tests. The issue of L2 gains made by international students while studying at university has been the subject of comparatively fewer studies (Knoch, Rouhshad, Oon, & Storch, 2015), even though this is an issue of substantial relevance to test developers and policy makers alike. It is not uncommon for policy makers to expect international students to make language gains by virtue of attending class in the L2 (Storch, 2009), even though recent research has contested the assumption that rich input alone will yield language gains (e.g., Gass, 2003; but see also Krashen, 1985) For language development to take place, learners need sufficient opportunities to use the language in a meaningful context (Ellis, 2003).

Offering support to the idea that language gains are not simply a matter of adequate input generating adequate output, recent studies have shown that even study abroad programs, which were designed to stimulate language gains, yield less demonstrable effects than educators and policy makers may want to believe (Sanz, 2014). Gains in oral fluency and lexical complexity have regularly been reported, but the findings for pronunciation, spoken accuracy, or writing are more ambivalent (Collentine & Freed, 2004; Díaz-campos, 2004; Hernández, 2010; Llanes, Tragant, & Serrano, 2012; Pérez-Vidal, 2014; Pérez Vidal & Juan-Garau, 2014; Serrano, Tragant, & Llanes, 2012).

Typically, study abroad research considers situations in which L2 learners go abroad for the purpose of learning a language (Engle & Engle, 2003). For most international students, however, developing their L2 proficiency is not the goal of their stay; the L2 is simply the main medium of instruction. Studies measuring language gains in this context exist, but usually focus on or include participants who received additional language support (Elder & O'Loughlin, 2003; Green, 2004; Llanes et al., 2012; Serrano et al., 2012). Only a handful of studies have examined language progress made by international L2 students who have not received additional formal L2 instruction, which is the default situation in many countries (Knoch et al., 2015).

Which language gains can be expected in international L2 students?

Longitudinal research into the language development of international students who do not receive additional language support is rather scant, but findings suggest that the gains are less substantial than commonly held beliefs would allow. Focusing on the language gains made by 24 Spanish undergraduates during one semester at an English university, Llanes et al. (2012) reported significant gains and medium to small effect sizes for written fluency (Words/T-unit: p = .032, r = .24) and oral fluency (Pruned syllables/minute: p = .000, r = .40). In a similar design, Serrano et al. (2012) traced language gains of fourteen Spanish undergraduates on an Erasmus exchange program in the UK. Statistically significant gains and medium to large effect sizes were reported for oral fluency (Syllables/minute: p = .004, d = 1.08), lexical richness (Guiraud's Index: p = .46, d = .65) and oral accuracy (Errors/T-unit: p = .011, d = -1.33). In the written modality, gains were reported in terms of fluency (Words/T-unit: p = .009, d = 1.11), syntactic complexity (Clauses/T-unit: p = .046, d = .83), lexical richness (Guiraud's Index: p = .41, d = .58), and accuracy (Errors/T-unit: p = .03, d = -1.15). Importantly, however, in both studies, a large proportion of the participants had voluntarily attended additional language classes (Llanes et al. N = 18/24; Serrano et al. N = 8/12), so it is impossible to determine whether these gains would have occurred had formal language support been absent.

Perhaps the only longitudinal research focusing on language gains made by international L2 students who did not attend formal L2 language support was conducted at the University of Melbourne, Australia. In a number of publications, researchers used a test-retest design to measure writing gains after one semester, one year, and three years. Storch (2009) asked 25 Asian international students to write a 300-word argumentative essay at the beginning of the semester, and again twelve weeks later. She found no statistically significant gains on fluency, accuracy, or grammatical and lexical complexity. Arguing that one semester might be too short a time-frame to observe language gains (see Ortega, 2003), Knoch, Rouhshad, & Storch (2014) then used the *Diagnostic English Language Assessment* (DELA) to measure writing gains in 101 participants over the course of one academic year. They found no gains in writing scores, no gains in terms of accuracy and complexity. Only fluency, as measured by the amount of words written in the time allowed, significantly increased, with a large effect size (p < .004, η_p^2 = .216). Lastly, covering a three-year data collection period that included 31 participants, Knoch et al. (2015) reported similar results: no gains on the DELA, and no gains in terms of discourse measures, written fluency excepted. After three years, participants wrote significantly (p < .003) more words within the 30-minute time frame.

Little if any research has investigated spoken language gains made by international L2 students who received no additional language instruction.

Identifying causes for limited gains

Only a few studies have examined the causes for limited language gains made by international L2 students. In the study abroad literature, authors have pointed to individual and contextual parameters as possible explanatory variables. In the study by Llanes et al. (2012), participants with positive attitudes towards the experience abroad, and participants who interacted more with L1 users made more gains. Dewey et al. (2014) reported that students enrolled in programs that stimulated interaction, made more gains, and that – in contradiction to some earlier studies – older students made more gains than their younger peers. According to Engle & Engle (2003) contextual variables too may impact language gains: longer stays, interactive didactic approaches, accommodation shared with L1 speakers, and opportunities for interaction with L1 speakers were argued to positively affect language gains. Here too, however, findings are contradictory. Housing conditions, for example, have been found to correlate with written accuracy (Llanes et al., 2012), but zero effects have also been reported (Magnan & Back, 2007). Similarly, in certain studies the length of a stay abroad impacted the gains made (Pérez-Vidal, 2014), while in others it did not (Elder & O'Loughlin, 2003; Serrano et al., 2012).

Again, research that offers explanations for limited language gains in international students is rather scant, but the studies that do exist typically suggest that one would have to consider the amount of meaningful interaction with speakers of the target language. A consistent finding in studies on L2 socialization, however, is that meaningful interaction with the L1 community is problematic, or rare (Ranta & Meckelborg, 2013). Gu & Maley (2008) identified four main areas in which the participants experienced problems adapting to the academic community in the UK: the culture of the host society, the didactic approaches, the language, and their new social role. In some cases, this resulted in boredom, loneliness and alienation. In a study that traced the academic socialization of six Japanese students at a Canadian university, Morita (2004) argued that international students' in-class participation does not necessarily depend on their intellectual abilities, but on the social roles they assume, or are allowed to assume. Kormos et al., (2014), focusing on interactions between international students and English students at universities in the UK, showed how L2 learners perceived negative reactions of L1 speakers to their language use as threatening, resulting in reduced willingness to seek out further contact with L1 peers. A study by Duff (2002) in a Canadian high school setting explained how participating in class allows L2 learners to show and develop their academic identity, yet the participants she described were caught in a dilemma that obstructed access to and acceptance in the discourse community: fearing that they would be ridiculed because of their English, the L2 learners in Duff's study

were afraid to speak, but also feared the other students' contempt when they kept quiet.

In summary, there is general consensus that international L2 students may find it hard to become legitimate members of a new academic community of practice for power-related cultural or linguistic reasons (Kormos et al., 2014; Ranta & Meckelborg, 2013; Subtirelu, 2014). Additionally, L2 learners who remain on the fringes of a desired community of practice have fewer opportunities for meaningful interaction, which is a key prerequisite for L2 learning (Hernández, 2010; Ortega, 2008).

Explanatory theories

A few current theories of L2 learning offer frameworks to explain L2 learning by referring to the social environment, including institutional and interpersonal dynamics, as mediating factors (Duff, 2002; Kinginger, 2004; Lantolf & Genung, 2003; Norton, 2013; Norton & Toohey, 2011, but see also Holmes, Marra, & Vine, 2011). In these theories, identity and social acceptance are of pivotal importance.

Influenced by Bourdieu's poststructuralist writings (Bourdieu, 1991), identity theory posits that identities shape and are shaped by the unequal social structures of the various communities in which we live (Norton & Toohey, 2011). Identity, in other words is defined by who we identify with, and by who we feel distanced from (Baynham, 2006). Thus, a sense of familiarity causes people to identify with one another (Kormos et al., 2014) and may lead to a cyclical reaffirmation of power structures. Norton (2013: 47) defines power as "the socially constructed relations among individuals, institutions and communities through which symbolic and material resources in a society are produced, distributed and validated". Similarly, Lantolf & Genung (2003: 178), conceptualize social power as the ability to see the world from one's own perspective, without needing to consider the perspective of others. As such, socially powerful groups dictate the terms of legitimate membership to their group or their community of practice (Block, 2007). Becoming part of such a group thus requires identity reconstruction and adjustment to new rules and norms (Lave & Wenger, 1992; Norton & Toohey, 2011; Swain & Deters, 2007). Since language learning is conditional on meaningful interaction (Amuzie & Winke, 2009; Ranta & Meckelborg, 2013), membership to these communities of practice is crucial for making language gains.

Research in the field of psychology supports the importance assigned to social belonging in these language socialization theories. Psychologists consider social belonging to be a basic human need (Baumeister & Leary, 1995; Walton & Cohen, 2007). In their extensive review of available research, MacDonald & Leary (2005) even argue that the human brain perceives social exclusion in much the same way as physical pain. In a similar vein, Charles Taylor has argued that our

identity is at least partly shaped by recognition from institutions and people around us, and in a thought-provoking essay he argues that non-recognition by others can cause actual damage, and may lead people to internalize depreciatory attitudes towards themselves (Taylor, 1992). Psychology has shown how this damage can manifest itself: feelings of exclusion may result in withdrawal, loneliness, sadness, or shame (Buckley, Winkel, & Leary, 2004). In a university context this may impact cognitive processing (Baumeister & Leary, 1995) and academic results (de Beer, Smith, & Jansen, 2009; Maestas, Vaquera, & Zehr, 2007; Walton & Cohen, 2007). Importantly, members of minority groups, such as L2 learners, may be disproportionally impacted by events that confirm a perceived absence of social connection (Walton & Cohen, 2007).

Recently, the Douglas Fir Group (2016) has proposed a promising and rather comprehensive framework that could advance research into language gains made by international L2 students. In their framework they combine interdisciplinary insights to distinguish three interconnected levels that could explain why language learning is obstructed or facilitated. In this chapter I will address the micro and meso levels directly (interpersonal, institutional), extrapolate to the macro level (ideological), and explain how the levels are interconnected.

The macro level of the Douglas Fir (2016) framework refers to matters of ideology. As such, it is closely connected to power structures (Subtirelu, 2014) and refers to the collection of ideas and values that speakers of a dominant discourse community commonly hold about language and the hierarchies between languages (De Costa, 2010, 2011; Subtirelu, 2014). When the dominant ideology is a monolingual one, negative attitudes towards speakers of languages other than the official L1 may ensue (Subtirelu, 2014; The Douglas Fir Group, 2016), especially when that language is perceived to have lower social status (Jordens, 2016).

Ideologies may remain at the level of language beliefs, where they shape and express latent consensus about what constitutes a good member of a discourse community (Spolsky, 2004), but they may also be translated into policy measures – the meso level. In many contexts, language ideologies implicitly or explicitly shape institutional policies (De Costa, 2010; Linton, 2009; Shohamy, 2006; Van Splunder, 2015). Studies have shown that a monolingual language ideology has become pervasive in many European nations (Gogolin, 2002), and the US (Linton, 2009). This may put L2 learners at a power disadvantage, since it rationalizes or reinforces the belief that some languages are dominant or superior (Subtirelu, 2014). Matters of ideology and power impact not only the norms of institutions and societies, but also the micro level of interpersonal interactions within them. They can influence the expectations towards L2 learners, the roles they can expect to assume, and the ideas L2 learners have about themselves (Douglas Fir Group, 2016).

The Douglas Fir framework can be seen as a broad, three-tier interpretational framework to explain language gains made by international L2 students that takes into account ideology, institutional dynamics, and interpersonal relationships. Typically, studies on international L2 learners focus on interactions with other students (Ranta & Meckelborg, 2013), on contact with ESL teachers (Lee, 2008), or on the academic culture of the host institution (Braine, 2002; Seloni, 2012). No studies in this field have yet proposed a broad explanation for language gains by linking institutional and interpersonal variables, and by considering them in relation to matters of ideology.

Summary of the existing research

It is clear that any presumption that international L2 students will make language gains simply by virtue of attending class in the L2, is unsupported by research. There is wide consensus that meaningful interaction with the L2 community is a catalyst for language gains, but in many contexts, this interaction appears rather limited.

Longitudinal studies that focused explicitly on language gains by international L2 students who received no additional language support, are very limited in number however, and have so far only considered writing gains. Additionally, no existing studies have combined an attention for oral and written language gains with the personal and academic experiences of the participants. Last but not least, apart from the Douglas Fir framework, very few comprehensive explanatory frameworks exist.

RESEARCH QUESTIONS

This chapter reports on the experiences of 21 international students at Flemish universities over an eight-month period. Crucially, Dutch is not a major world language, but a comparatively small language, which is rather exceptional in an English-dominated research field. As a language with twenty-two million native speakers, Dutch is of modest international importance. This makes the educational setting in which this study is carried out unique in at least two ways. One, it means that most international students have had little exposure to the language as medium of instruction, Dutch, before studying it. Secondly, for most international students Dutch is a medium to reach a desired outcome, but not the actual goal of their stay in Flanders.

The first research question considers language gains made by full-time students enrolled in five Flemish institutions of higher education who have an international L2 background. Gains were measured, perhaps for the first time in a study of this kind, by re-administering sections of the actual university entrance

test the participants had taken upon enrolment and by analyzing not only test scores but also measures of complexity, accuracy and fluency of their output. Importantly, the participants received no formal language support between the test and the retest. The performances on the test and the retest were inspected quantitatively and supplemented with the participants' perceptions for the purpose of triangulation to answer the first research question:

RQ1 *Did eight months at university lead to Dutch language gains among the L2 participants?*

Based on the results of previous research, it was hypothesized that participants would make oral language gains, but that the gains in the written modality would be limited at most.

The second research question, which is at the heart of the current study, focuses on explaining the outcomes by analyzing the experiences and perspectives of international L2 students during their first year at a Flemish university. Using qualitative interview data, the dynamics that exist within the university will be illuminated at two levels, the institutional and the interpersonal:

RQ2$_a$ *At the interpersonal level, how might student-student and student-teaching staff interactions at Flemish universities impact the language learning opportunities of international L2 students and hence their chances at making linguistic gains?*

RQ2$_b$ *At the institutional level, how might the perceived positions of L2 learners at their university impact international L2 students' language learning opportunities?*

By operationalizing the Douglas Fir framework, this study was designed to identify interactions between institutional and interpersonal variables that may inhibit or promote language learning.

Additionally, this study generated data about drop-out. During or at the end of the data collection period, some participants left university. The third research question links their reasons to do so to institutional and interpersonal variables:

RQ3 *To what extent could institutional variables or interpersonal relationships explain certain participants' decision to leave university?*

This study relies on a sequential explanatory design, in which longitudinal qualitative data serve to interpret quantitatively measured language gains

(Creswell, 2015). The measurement of language gains draws on a repeated-measure within-subject methodology (Kinginger, 2008; Knoch et al., 2015, 2014; Storch, 2009). The results pertaining to the other research questions primarily draw on patterns identified in the analysis of the coded interview transcripts.

PARTICIPANTS & METHODOLOGY

Participants

The twenty $L2_F$ participants who participated in this study were selected from the larger $L2_F$ population. This group was already introduced in Chapter 2, but certain characteristics relevant to this chapter will be reiterated or highlighted here.

To be eligible for participation in the longitudinal leg of the research project, a number of inclusion criteria had to be met. Participants had to register for university – the academic year following the language test. PhD students were not considered because they are not required to attend classes and because they do not need to meet language requirements; neither were students who signed up for an English-medium program. Out of thirty-two possible $L2_F$ participants, twenty agreed to be participants. Appendix 4 lists a number of demographic variables per participant, and includes information about their study success in July 2015. A "+" indicates that the participant passed more than half of the courses he or she had registered for. This cut-off point was based on the study success of international students in Flanders, who (on average) attain 47.7% of the credits they take on (Glorieux, Laurijssen, & Sobczyk, 2015: 15).

The median age was 23 for the ten bachelor students (min = 19, max = 32) and 24 for the ten master students (min = 19, max = 44). The overall mean age in Flanders is 21 for bachelor students, and 24 for masters (Wartenbergh et al., 2009). French and Spanish were the most predominant L1s, and most participants were from Europe and Latin America. Four participants were from the francophone southern part of Belgium, where the regional government sets out an educational policy that is independent from the Flemish one. These twenty participants registered for a program within one of the main academic traditions: humanities, exact sciences, and social sciences. Importantly, Dutch was their medium of education, but Dutch language learning was not a goal in itself – except for Oksana who studied translation studies with Dutch as the target language. Most participants were enrolled at one of the three largest universities in Flanders: Ghent University (6), the University of Leuven (8), and the University of Antwerp (4). Leila attended an interuniversity program that included these three universities. Stella had intended to register at Ghent University, but was denied entry because she had not passed either university

entrance language test. She then registered at the smaller University College of Hasselt after passing the local test.

Thirteen participants took part in the project for its entire duration. Such attrition is to be expected in longitudinal projects, even more so in view of the fact that university education in Flanders has large dropout rates. Forty percent of the students drop out prior to the July examinations of their first year at university (Goovaerts, 2012). In this study attrition was not necessarily regarded as a loss of data however, since the reasons why participants discontinued their studies was highly relevant in the light of the research questions.

The participants received no remuneration for their participation. When asked in June 2015 why they had chosen to remain in the project for so long, the reasons they stated were being able to talk to somebody (e.g., Gabriela, Oksana), feeling part of a group (e.g., Elena, Alexandra), practicing Dutch (e.g., Marie, Guadalupe), or doing something that matters (e.g., Merveille).

Data collection: test-retest, monthly interviews, and field notes

The data collection for this study lasted from July 2014 until June 2015, and included an initial language test, semi-structured interviews, and a retest, which included a listening-into-writing task and an oral presentation task.

Test/Retest design

All participants had taken STRT in the summer of 2014. For the retest in April 2015, it was decided not to run STRT fully, but to select two representative tasks. This was done mainly for reasons of time; STRT takes up to four hours, and at a time when the final exams were coming up, it was not considered in their best interest to ask the participants to devote this much time on a retest. Based on data from the previous full-scale STRT administration that included the same tasks (N = 913) a multiple linear regression was run in *R* (*QuantPsyc* and *car* packages) to determine the oral and written tasks that explained most score variance (see Table 7.1).

Table 7.1. Multivariate linear regression: STRT total score ~ T1-T6 scores

	B(SE)	*β*
(Intercept)	-0.008 (.011)	
T1	1.396 (.001)***	.161
T2	1.451 (.001)***	.257
T3	1.628 (.001)***	.167
T4	1.397 (.001)***	.216
T5	1.860 (.002)***	.161
T6	2.268 (.001)***	.295

Note. R^2 Adjusted = 1, $p < .000$

Based on the standardized beta values, the oral presentation task (β = .52) and the writing-from-listening summary task (β = .57) were selected for the retest. Together, they explained 90.8% of the overall STRT score variance (R^2_{adj} = .908, p < .000). Appendix 1 offers more details concerning these tasks.

Monthly interviews

The participants in this study were interviewed every month of the academic year, except during the study-intensive months (December, January, May). The interviews were semi-structured and typically lasted around 45 minutes (Median = 44, Min = 24, Max = 97). Every month the interviews had a different focus, but issues regarding the participants' social life, academic experiences and perceived language progress were always on the agenda. Figure 1 shows the timeline of the study and indicates which participants were no longer involved at which point.

In order not to create a further disequilibrium in the power balance that is inherent to interviews, the participants always had a say in determining the time and place of the interview (Sin, 2003). Some preferred a café, some wanted to meet in the university cafeteria, and others felt comfortable to meet at my workplace, where the seating had been arranged to create a collaborative atmosphere. The researcher invested energy in developing a sense of mutual understanding or communality by sharing personal details when requested, and by establishing rapport (Manderson, Bennett, & Andajani-Sutjahjo, 2006). Familiarity with the interviewer can generate a sense of comfort and security (Pellegrino Aveni, 2005), and the participants in this study felt free to discuss personal matters. At the end of the year, some participants (e.g. Alexandra, Guadalupe, Oksana, Gabriela, Julia, Hoang, and Anastasia) acknowledged that they had shared thoughts and experiences with me that they had not shared with other people they trusted.

Field notes

The researcher kept field notes during the data collection period. In this chapter, only the notes taken in November 2014 will be referred to. During that month the researcher accompanied the participants to a class of their choosing, taking note of the interactions and of the participants' responses to the in-class dynamics. All classes were audio-recorded, except for one, because the professor had not given permission to do so.

Analysis

Test data

The performances were anonymously double rated by two independent trained STRT raters who used the STRT rating scale. The STRT rating procedure takes into account content criteria and linguistic criteria, which are scored on an ordinal four-point scale. Due to the ordinal nature of the scores and the relatively low number of participants, Wilcoxon's Signed Rank Test was used to determine the significance of score differences (Field, Miles, & Field, 2012), and a function was created in *R* to calculate the effect size *r*.

Since it is possible that band scores are too broad to capture language gains over a period of a few months (Green, 2004), gains in term of complexity, accuracy, and fluency were calculated too, using a methodology based on Llanes et al. (2012) and Serrano et al. (2012). Written and oral fluency were determined respectively by computing the amount of words per T-unit, and the amount of syllables per minute (i.e., pruned to exclude repetitions, false starts and the like). Lexical complexity was calculated using Guiraud's Index. The measure for accuracy was the amount of errors per T-Unit for the written data and the proportion of errors per AS unit for the oral data (Foster, Tonkyn, & Wigglesworth, 2000: 365). Wilcoxon's Signed Rank Test and effect sizes were used to determine the significance and magnitude of possible gains.

Interviews

All transcriptions were coded and analyzed in NVivo 11 For Mac. Based on the literature review, an a priori coding scheme with fixed coding categories was set up (Miles, Huberman, & Saldaña, 2013). The main branches of the a priori coding tree were "Background variables" (6 categories), "Academic work" (11 subcategories), "Language use" (7 subcategories), "Language tests", and "Identity" (20 subcategories). I coded all interviews using these categories but was free to add an open, inductive layer of coding, when salient themes emerged (De Costa, 2011). One such coding was "Key transformational episode", which was used to mark instances or anecdotes that had a profound impact on a participant and recurred in different interviews with the same person. To check for coding accuracy a trained research assistant coded the November 2014 interviews (54239 words) using the a priori categories. There was a an exact agreement between the raters of > 90% (Landis & Koch, 1977).

After data analysis, all coding categories relevant to the research questions at hand were combined into data matrixes (O'Cathain, Murphy, & Nicholl, 2008), which included the most prominent indicators of variables affecting language gains on the institutional and interpersonal level (see Dewey et al., 2014; Engle & Engle, 2003). On the institutional level, four variables were included: classes, examinations, use of support systems, and language use at university. The analysis of the interpersonal level focused on social networks at university (i.e., teaching staff and students). As an addition to manual coding, the most frequently occurring content words in the transcriptions of the participants' speech were identified. These text searches offer a complementary perspective on the data.

RESULTS

Linguistic gains after eight months at university

Between the first STRT administration in the summer of 2014 and the second, in April 2015, no significant speaking or writing gains were made, as measured by STRT. The median score (see Table 7.2) indicates a slight non-significant score gain on the written summary, with a medium effect size (r = -.31). The median score for the presentation task has decreased over time, but the effect is negligible (r = -.05) and the difference is not significant.

Table 7.2 Language gains (N = 13)

	Median Test	Median retest	*p*	*r*
Written summary				
STRT score	17.25	19	.159	-.31
N words	207	240	.07	-.4
Lex complexity	56.5	57.5	.89	-.03
Synt complexity	.26	.25	.79	-.06
Accuracy	.94	.80	.8	-.06
Fluency	9.9	11	.85	-.04
Presentation				
STRT score	28	27	.824	-.05
N words	424	345	.03	-.5
Lex complexity	38	37	.96	-.01
Synt complexity	.41	.38	.48	-.17
Accuracy	.86	.86	.66	-.1
Fluency	12.1	13.7	.73	-.08

The only significant ($p < .05$) difference and the largest effect (r = -.5) is the decreased amount of words used in the retest of the presentation task. The second largest effect (r = -.4) is the increased number of words in the written summary. Word counts convey no information about the quality of a performance however. All indicators of text quality – syntactic and lexical complexity, accuracy and fluency – showed that no significant or substantial language gains were made. In terms of effect size, the largest difference is for oral syntactic complexity, which decreased slightly between the test and the retest (r = -.17).

The quantitative results were confirmed by the overall intuitive estimation of the participants. At the end of the year, no participant felt that his or her overall Dutch language ability had improved. Some felt more self-confident when using Dutch (Guadalupe), or believed that one or two skills had improved: A few participants reported perceived gains in writing (Marie), listening (Emma, Merveille) or lexical range (Marie, Ersi). On the other hand, some also believed that their speaking ability had decreased (Ersi), or that no gains whatsoever had been made (Oksana, Elena, Gabriela).

Interpersonal relationships

Table 7.3 shows the recurrence of the words "Different" and "Difficult" in the interviewee transcripts. A detailed analysis of the interviews confirms that all participants had a difficult time adjusting their new role in a different reality. This remained true at the level of all interpersonal relations.

Table 7.3. Reduced data matrix: interpersonal

	Codes	Interviews	Frequent words	Times used/code
Faculty staff	35	25	Different	.8
			Dutch	.7
			Question	.6
Students	437	65	People	1
			Different	.86
			Difficult	.50

Table 7.4 shows an overview of the perceived attitude of the participants' teaching staff and their L1 peers, and indicates what the participants considered their best and their worst social experience of the year. The interview analyses will be discussed in detail below, but the trends are clear from the table: only one participant described the teaching staff as involved, while others mostly described their professors and teaching assistants as distant, with occasional positive (respectful, kind) or negative (demotivating, face-threatening) qualifiers. Most participants also described the attitude of their L1 peers as closed or distant. In the first semester only Elif described the attitude of some of her fellow students in a positive way. After one year, not much had changed. There still was no regular contact between the participants and their L1 peers.

Table 7.4. Interpersonal relationships

	Group size	Perceived attitude of teaching staff	Perceived attitude of L1 students		Social acceptance at university (June)	
			October	June	Worst moment of year	Best moment of year
Elena	100+	Distant, formal, kind	Closed	Closed, xenophobic	Consistently low	No best moment
Alexandra	100+	Distant	Closed	Closed, ignoring	Ignored by classmates	Chit-chat after exam
Marie	300+	Distant, kind	Closed, hard-working	Distant	Ignored by classmates	Occasional bar visit
Leila	50+	Distant	Closed, face-threatening	Closed	Consistently low	No best moment
Guadalupe	500+	Distant, neutral	Closed, childish	Closed	Ignored by classmates	Studying with friend (L2)
Elif	50+	Distant, respectful	Some are welcoming	Distant	Consistently low	No best moment
Oksana	50+	Sometimes demotivating	Closed	Closed	Consistently low	No best moment
Ersi	500+	Distant, respectful	*No contact (illness)*	Friendly, hard-working	No lows	Going to conference
Alireza	300+	Distant, some disrespectful	Closed	Closed, superficial	Consistently low	No best moment
Merveille	300+	Distant, respectful	Closed	Closed	Ignored by classmates	Day trip with housemates (L2)
Gabriela	500+	Distant, impersonal	Closed, superficial	Closed	Ignored by classmates	Attending wedding (L2)
Emma	500+	Inapproachable	Closed	Increased openness	Consistently low	No best moment
Hoang	500+	Distant, some are kind	Closed	Increased openness	Ignored by classmates	Afternoon of group work
Anastasia	50+	Inapproachable	Closed	*Left university in February, stated interpersonal relationships as one of the reasons*		
Océane	500+	Distant, formal, kind	Closed	*Left university in February, stated interpersonal relationships as one of the reasons*		
Clara	500+	Distant, face-threatening	Closed, hard-working	*Left university in February, stated interpersonal relationships as one of the reasons*		
Stella	50+	Involved	Distant, friendly	*Forced return to Armenia in March*		
Yazdan	500+	Distant, neutral	Closed	*Left university in November, allowance reduced*		
Jessica	50+	Distant	Treated with suspicion	*Left project in November, reason unknown*		
Chloé	500+	Distant	Closed	*Left project in November, reason unknown*		

If we inspect the evidence about students' perceptions of the teaching staff at their university, the relationship can best be described as distant. Because of the predominant teaching style and because of the group sizes, which could be as large as 500 enrolled students in a classroom, interaction between the participants and the professors was mostly non-existent, or limited to an occasional question. Stella was the only participant who reported regular, meaningful interaction with her teachers. The nineteen other participants who were asked about their perception of the teaching staff during the first interview described contact with teaching staff as limited, also when they were enrolled in smaller programs of about 50 students.

For three participants the few encounters with their professors had been positive experiences, but for others these interactions confirmed the feeling of hierarchic distance. The only participant who described her professors as involved, was Stella. Yazdan, Merveille en Elif described the relationship with their professors as distant, because professors had too many students and too little time, but overall positive. Other participants reported not interacting with professors (e.g., Emma, Elena), or shared anecdotes of interactions with professors that were perceived as condescending, distance-inducing or even threatening (Alireza, Leila, Clara, Anastasia).

Marie, a francophone student in Flanders claimed that her professors reacted rather positively to her being an L2 learner of Dutch. Others, however, offered frequent anecdotes of negative encounters with professors concerning language-related issues. At least nine participants reported feeling uneasy about being an L2 student in class, but not all of these participants had concrete anecdotes to substantiate this feeling. Clara discussed concrete episodes, such as this one:

> Last time the professor asked me a question and I got all red and I said "sorry I speak French, I didn't understand everything". She looked at me without a smile and then she looked into the auditorium and said "madam speaks French".
>
> (Clara, November 2014)

Likewise, Oksana, a Ukranian-born student of translation studies, described how one professor had reacted quite negatively to her attending class as an L2 learner:

> So when I needed to discuss my curriculum I emailed one of the professors, who said "I'm sorry but if you make such basic mistakes in an e-mail, this program will not be attainable for you".
>
> (Oksana, October 2014)

One month later, I accompanied Oksana to a class taught by the same professor and took note of the following interaction.

> The professor asks her a question about a word in the dialect of coastal Flanders.
> "You are *otherlingual*, right? What's your mother tongue?"
> "Ukranian and Russian"
> "I bet they don't say *this* in Ukranian" [pronounces dialect word, classroom laughs]
> [...]
> After class, he comes over to Oksana and me. They have never spoken before. He asks her which courses she has and how long she has been studying Dutch. "Six months", she says. [...] He is surprised at how good her Dutch is, but tells her that she will probably not pass, that she should consider quitting and finding a job.
>
> (Field notes November 2014, Oksana)

The negative nature of this interaction is an exception in the field notes. In all other classes observed, with the exception of one (Stella), professors would typically deliver an ex cathedra talk. Some did not interact at all, as in Alexandra's class:

> There's a lot of loud talking while the professor keeps on talking without either asking for silence or stimulating dialogue.
>
> (Field notes Alexandra, November 2014)

Other professors who were observed for the current study kept a distance and made occasional comments without really interacting, but most delivered a two- to three-hour monologue in an auditorium. The setting of Anastasia's class was exceptionally uncomfortable:

> "The room is very dark and it's hard to see my notes. The benches are hard. Teaching assistant delivers monotonous monologue sitting down, power point projected above her head. [...] At one point a student whispers something. The TA immediately stops talking and says *excuse me, this is very bad for my focus*".
>
> (Field notes Anastasia, November 2014).

All participants confirmed that the class I had attended was representative of other classes of the same course.

In short, most $L2_F$ participants perceived their relationship with the teaching staff as distant. This was also the general perception of their relationship with Flemish L1 classmates. With the exception of Ersi, all participants found it difficult to befriend Flemish students. From the analysis of the data, three main patterns can be identified: (a) networks among Flemish students appeared impenetrable to the participants, (b) many participants perceived a power imbalance between them and the Flemish L1 students, and (c) many participants had to renegotiate their academic identity. In what follows I discuss each in turn.

The theme of closed L1 networks was prominent during the whole year. With 227 instances, *Social distance from peers* was the most frequently used code in all transcriptions combined. Most participants agreed that Flemish students made a closed impression, and all were surprised by the difficulty of befriending Flemish students. In the October interview fifteen out of twenty participants mentioned feeling stupid, incompetent or unlikeable among their L1 peers. For students like Anastasia, the perceived impenetrability of L1 networks and the perceived indifference of her L1 peers were the final argument in her decision to drop out.

> After the first month it was clear that I would not be happy here. The students pretend like I'm not there. I don't think they show any interest. They talk among each other and when I say something, they answer and start talking to each other again.
>
> (Anastasia, February 2015)

Particpants who registered for master degrees referred to already-formed networks of students which were hard to penetrate. But also first-year students like Gabriela, Hoang, and Chloé did not gain access to L1 student networks. To the participants these semi-impenetrable groups seemed to have been formed in secondary education, or appeared to be regionally determined.

> Some Flemish people are a little difficult to befriend I think. […] I talked about this with other international people and everybody says the same things: you guys have got the same friends from when you're like three till you're like seventy. It's very hard to find your way into a group of friends.
>
> (Gabriela, October 2014)

> Somebody else told me "we already have a group of friends, why add a francophone?" He wasn't being mean, but it was like "we are from Hasselt", "we are from Kortrijk" [two small Flemish cities with distinct dialects].
>
> (Chloé, October 2014)

In June, when asked which aspect of the last year they would have liked to change, six out of the remaining thirteen participants referred to the disposition of Flemish students, which was characterized as polite and friendly but distant. For a few participants, the situation improved in the second semester, as they became acquainted with Flemish students. But even so, only Ersi and Alexandra reported having L1 friends in class. Others (Marie, Merveille, Hoang and Emma) did not go as far as to call any L1 classmates their friends.

Most participants remained distant from their L1 peers, but not all were affected by this in the same way. Alireza, rather introvert by nature, did not really mind the lack of contact. Leila and Guadalupe, respectively 44 and 30 at the time of data collection, also felt distant from their fellow students, but did seek out contact because of the age difference. Nevertheless, irrespective of how the participants responded to the attitude of Flemish students, all of them perceived it as closed.

> *- Can you describe the contact between you and your fellow students from Flanders?*
> Non-existent. Nothing.
> *- Do you mind that situation?*
> I used to. Now? No. I have lots of friends from the Ukraine, from Russia, from everywhere. I have friends I can go to bars with. I am not alone.
>
> (Elena, April 2015)

Elif and Ersi were the only participants who had made Flemish L1 friends within their study program during their first semester as students. Ersi had made a few friends during a conference that she and her classmates had attended. During the first semester Elif, who was enrolled in a comparatively small and specialized master's program, had occasional dinners with a few classmates, some of whom were Flemish.

That does not mean that all Flemish students were closed however, and not all participants were socially isolated. Quite a few had a group of friends outside university, and although most participants usually interacted with other international students, some had Flemish friends outside university. Merveille, for example, had quite a few friends at her student accommodation. Anastasia, Oksana and Alexandra had Flemish boyfriends at the time of data collection, but they did not consider their boyfriends' friends as their own. Their descriptions of New Year's Eve for example were strikingly similar.

> I didn't want to go out with my boyfriend's friends because they aren't my friends. [...] So I studied until 11.30PM and that was really hard. But the next day I bought a ticket to Peru.
>
> (Alexandra, February 2015)

> I just went out to celebrate with my boyfriend's friends. But I didn't like it. I didn't feel like celebrating. I'm having a hard time here without my friends. […] And I haven't seen my family in two years.
>
> (Oksana, February 2015)

Loneliness, in sum, was a pervasive feeling among the international students in reaction to the closed L1 networks:

> My biggest problem here was loneliness. I have almost no contact with my classmates. Sometimes I just have questions or I don't understand something, and then there is nobody to help.
>
> (Anastasia, February 2015)

But feeling excluded from L1 friendships and coping with loneliness was not the only challenge the international students faced. Throughout the year, participants referred to instances of power imbalance, which in most cases resulted from feeling inferior in terms of language ability, not quite understanding the culture, feeling excluded socially, or fearing ridicule. In March, all participants answered the question "How do you think the other students see you?" None of the answers given (see Table 7.5) indicate that their fellow students know their personality. Two participants offered a somewhat positive response, but they too focused on differences rather than communalities. In quite a few cases, participants shared stories of exclusion. Typically, these stories revolved around one participant wishing to be included in a network of L1 peers, but not gaining admittance. In March, Gabriela recounted one anecdote she returned to during every subsequent interview:

> This has happened like two or three times. I'm sitting here, right? And there's two girls here [points right] and here [points left]. I talk to these girls sometimes, or occasionally, or whatever. Anyway so one girl asks to the other "do you want to get a coffee"? And I'm right in the middle! And then I wonder can I go too but then I think I'm in the middle and they don't ask me. These typical things I don't understand. I think it's terrible.
>
> (Gabriela, March 2015)

Table 7.5. How do think your classmates see you? (March)

Negative	Elena	That I am stupid
	Guadalupe	That I am weird
	Alireza	They think of Iran in stereotypes
	Marie	They are surprised. Sometimes they think that we are lazy.
	Hoang	I hope they can see that I am kind.
Indifference	Alexandra	They are indifferent
	Leila	I am isolated
	Elif	I don't know
	Oksana	They don't pay attention to me
	Merveille	They just say hey and that's it
	Gabriela	They aren't interested
Positive	Ersi	They appreciate that I learned Dutch so quickly
	Emma	Some are surprised, some love that I am foreign

In some cases, the inhibitions and preconceptions of participants obstructed interaction with their L1 peers. For Hoang and Leila, language imbalance made them feel like an outsider. Both had been addressed and invited out by L1 students, but kept their distance because they felt linguistically inferior.

> I don't dare to speak because [the L1 students] talk so quickly and you don't understand. But then you try to say something and they don't understand. It makes things difficult. Such shame. I feel such shame when I need to speak to somebody.
>
> (Hoang, February 2015)

> I have little contact with the other students. In this class for example, this is an English course and we are equal. We are on equal terms when it comes to language, and I feel no embarrassment, no shame. I feel like an outsider, but they are outsiders too.
> - *Why do you feel like an outsider?*
> That's normal. They have an advantage that I lack. They are Dutch speakers in a Dutch course. I am francophone and my Dutch is not perfect. That is different. Sometimes I wanted to ask a question, but told myself "nonono".
>
> (Leila, July 2015)

Occasionally too participants asserted power over their L1 peers by stressing differences in age and perceived maturity in a derogatory way. For example, at different occasions participants referred to their fellow students as children, babies or teenagers (Leila, Guadalupe, Anastasia, Hoang, Alexandra), or pointed

out differences between themselves and their L1 peers in terms of financial means or parental support (Anastasia, Hoang, Alexandra).

By the end of the second semester, most participants had become milder in their opinion about the Flemish students' behavior. Quite a few (e.g., Hoang, Emma and Gabriela, Elena, Guadalupe) had accepted that there was a power imbalance, and that international L2 students had to make a larger effort to gain social acceptance than Flemish students:

> I understand. You live here and you get all these foreigners, and you close up. It's normal. I'd do the same. […] You can't expect Belgians to go like "come here, strange person".
>
> (Guadalupe, March 2015)

As can be inferred from the results above, most participants felt somewhat out of place and socially isolated at university. Over time, however, they may have started feeling more legitimate as peripheral participants in the world of the university. Halfway through the second semester the participants were asked whether they felt at home in class. Four participants confirmed yes, seven were in doubt, and two did not feel at home. These two students motivated their answer by referring to being out of place as a non-Flemish student at a Flemish university:

> The classes are for Flemish students […] Not much attention goes to non-Flemish students
>
> (Hoang, April 2015)

When discussing why they did or did not feel at home in class, the students mentioned three primary reasons: understanding the language of instruction, having friends, and understanding the content of the class. For Merveille and Guadalupe, for example, realizing that they had started to understand the classes, marked a clear transformation in how they perceived of themselves as students.

Clearly, feeling at home in class was linked to having friends there. In most cases the sense of loneliness may not have been the product of Flemish students willingly and noticeably underappreciating the international students, but sometimes Flemish students were perceived as not valuing the contributions of their international peers, or as sabotaging them academically. For example, Océane recalls how one Flemish student told her that his class notes were free for his Flemish friends, but that she had to pay €20. Elena received no help when she asked classmates to check a Dutch text she had written. And the Flemish members of one group Alexandra belonged to made agreements without involving her:

> I was involved in group work and we were supposed to speak English but they always spoke Dutch. So we speak and yeah when I had ideas they heard and ignored me in a certain way. I'm there, I always go on time, read everything – except maybe once – but they ignore me, do not include me, and change the topic.
>
> (Alexandra, November 2014)

Even though the situation for Alexandra improved during the year, and she gained more positive experiences with other groups, the feeling of being excluded from the academic community persisted and still impacted the way she felt about being part of the academic community at the end of the year:

> I went to the library [...] I saw some classmates, but they just ignored me. [...] That was really hard cause I really try. I try to keep my distance from people from my country. I try to get integrated, but it doesn't work. I'm not looking for a deep friendship, but at least some eye contact, you know? [...] I was really crying. I didn't want to go to the library anymore. I had expected maybe to see somebody or to talk to somebody or if I doubted something in the course, to find a solution with somebody.
>
> (Alexandra, June 2015)

Lastly, understanding course content appeared to be a precondition for the participants' feeling of being a legitimate member of class. Participants who had started to feel at home in class by the second semester include Guadalupe, who had started the year with relatively low expectations, but had started studying on a daily basis after attaining good grades in January. All participants who did not really feel at home in class and mentioned study-related reasons to substantiate that opinion, stated that they no longer were the student they remembered themselves to be. Elena, Alexandra, Guadalupe, Oksana and Anastasia talked about how the feeling of being an accomplished student had been replaced with a sense of doubt:

> Here I sometimes feel stupid. [...] I had a good job in the Ukraine. There was a competition that included professors and everything. But I got the job. I worked at the airport as an aviation engineer. [...] Here there was this situation in class when I didn't understand something as quickly as the others did, and some people laughed, like "hahaha, you are from the Ukraine and you are stupid".
>
> (Elena, February 2015)

> Once, we had a guest lecture in English and it dealt with linguistics and the lecturer asked many questions and I could answer, because it dealt with linguistics and it was in English. I knew everything! And for the first time I felt at home. Normally I am just confused all the time: "Why am I here? I don't understand anything."
>
> (Oksana, February 2015)

Institutional support

Feeling at home in the educational system

Table 7.6 shows that the most frequently used words in the participants' transcripts were "difficult" and "different". The interview analyses also show that no participants found their way into the system without a hiccup, but that most participants became accustomed to the organization of their university in the course of the second semester.

Table 7.6. Reduced data matrix: institution

	Codes	Interviews	Frequent words	Times used/code
Classes	286	65	More difficult	.68
			Different	.66
			Everything	.43
Examinations	149	43	Difficult	.82
			Passed	.60
			Different	.60
Support systems	107	47	Different	.55
			Dutch	.45
			Students	.42
Language use	277	59	Dutch	.76
			Difficult	.71
			Different	.48

Three aspects of the institutional organization of Flemish universities turned out to be hurdles to learning: the large class sizes, the teacher- and lecture-centered pedagogical approach, and exams. In what follows, each will be briefly described.

During every interview the participants were asked to characterize the didactic culture they experienced at university. Overall, the most commonly used characterizations were "not interactive", "big groups", "boring", "solitary", and "scary". The first two labels of this list appeared in every interview conducted in October, but as the year progresses, participants started using them to a lesser extent. In November, "boring" is used most often, and after February, "solitary" becomes the term most often used to characterize the classroom experience. These lists of terms indicate that the participants did not enjoy their classes at university, especially during the first semester.

After their first class, all participants mentioned the lack of interaction as a striking characteristic of the classroom pedagogy, but it was not necessarily seen as a disadvantage however. Most participants felt that an ex cathedra style of teaching made classes rather monotonous, but on the other hand, they were happy that it shielded them from having to speak in such big groups. Throughout the year, fifteen out of the twenty participants expressed language-related fears about answering questions in class. Group sizes had an impact on their willingness to speak in class, but also made it easier to disappear in the crowd.

> It's a bit scary, so I don't like to speak in class, because I am afraid to show myself.
>
> (Hoang, November 2014)

> I had class in big auditoria for 500 students [...] at first I was shocked, but now I have friends it's no problem. At first I was totally alone there. I felt so small.
>
> (Ersi, June 2015)

In October, only Stella, Merveille and Jessica shared positive comments about their in-class experience, but as the year progressed, the opinion of some participants started to shift. By March, three out of the remaining thirteen participants still did not feel at home in class (Hoang, Elena, Oksana), but four did (Ersi, Emma, Merveille, Alireza) , and the remaining six did so to a certain extent (Gabriela, Alexandra, Marie, Guadaloupe, Elif, Leila). Correspondingly, the number of negative words used to describe class decreased, and in July, at least six out of the remaining thirteen participants looked back on their classroom experiences in somewhat positive terms.

> It's a shame that I don't know many people, but apart from that it was fun. I learned a lot, and I liked it when I understood the professor. […] When I listen to the professor, I really feel like one of the other students.
>
> (Guadalupe, June 2015)

The style of teaching had not changed, but some participants had adapted to the didactic culture, which nevertheless remained a daunting experience for others:

> In hindsight, sometimes we were too many in one room. We were like 300 or more. […] We had some interaction in one course. The professor would pick two students to sit in front; and he would ask them questions.
>
> (Marie, June 2015)

Other participants never felt at home in class for the entire year.

> If I'm being totally honest, this year has been the worst experience I have ever had. […] I saw the professor talk with other students, and sometimes they were laughing, like "oh well done" or "maybe pick a different answer", and with me it was always like "no, that's wrong". It wasn't in a sad, or angry, or irritated way, but I never heard "well done". At one point I just wanted to hear that I had done something right. But never, and I thought "come on, I worked so hard", and all for no positive feedback.
>
> (Oksana, July 2015)

At Flemish universities there are three examination periods. One in January, after the first semester, one in May, after the second semester, and one in September, for students to retake exams they failed during the year. Typically, all exams are planned in a two to three-week period, which is preceded by a three-week study break.

During the first interview, in October, eight participants spontaneously brought up the topic of examinations. In every instance, the participant mentioned fearing the examinations because they felt that they did not have the language competence required to successfully sit a written or an oral exam.

> Speaking on an examination is something I can't imagine myself doing right now – all those difficult words I just don't know yet. […] Most exams are oral, some are written, but that is tricky too. Environmental law is a written exam. It has written questions you have to provide a written answer to, using difficult words that I don't even know how to write.
>
> (Anastasia, October 2014)

After the first examination period participants did not perceive language as the main problem they encountered however; loneliness, stress and monotony were the most frequently mentioned problems the participants associated with the January examinations. For some (e.g., Alexandra and Elif), the exams were a challenge they were proud to have withstood. For others, it had been a demotivating experience.

> You always just eat, study and sleep. That was boring for me. Super boring. […] I think I haven't seen many people in a month and was just alone at home. On some days I didn't do much because it was so boring. […]
>
> (Gabriela, February 2015)

When the January exams began, most participants were unprepared. Many were surprised by the amount of studying and by the memorization that is expected in many courses. Quite a few participants had been unable to study all the required material, and many felt that they had started studying too late, or that they did not know what was expected of them.

> Flemish students know what to study. "Like one of my classmates said why are you studying that? It's nice to know but not necessary to pass." […] They grew up with this implicit knowledge [of how you need to study].
>
> (Alireza, April 2015)

Marie, Oksana, Elif and Alexandra were the only participants who were happy with their results on the January examinations and who did not change their study approach in the second semester. All other remaining participants decided to study more, to study more systematically, and to focus more on memorization.

> The level is high here but when Belgian students study, they memorize. I expected questions like "explain this", but all they do is memorize.
>
> (Elena, June 2015)

After the July examinations, eight out of the remaining sixteen participants had passed at least half of the courses they had taken on.

Preparedness for the university language demands

Here, we briefly revisit the $L2_F$ participants' readiness for the real-life linguistic demands of academic studies at university that were discussed in Chapter 2, but from a longitudinal perspective, and with attention to the impact of language proficiency on identity.

At the start of the academic year, none of the participants felt fully prepared for the listening demands of university. They felt unprepared for the variety of accents used by lecturers, for the variety in terms of pronunciations, and for the lexical range they were expected to master. A few participants reported not having understood anything of the first classes they attended (e.g., Yazdan, Océane, Leila), and some students never quite managed to get over the initial shock they felt when they attended their first class and discovered they understood little or nothing. Emma, for example, described a feeling of panic when confronted with Dutch during the first semester. During the second semester this feeling faded away:

> I was really panicking in the beginning. Really panicking. I didn't see a way out. [...] It was just too much, and I understood so little.
>
> (Emma, November 2014)

> *- Can you name three key words that describe the examination period for you.*
> Just panic. [...] Panic. Yes. Because of the language problem.
>
> (Emma, February 2015)

> [My new study strategy] works better. I won't panic anymore if I don't understand something.
>
> (Emma, March 2015)

> I have no idea what happened during that first semester. I really don't know. I wasn't thinking, I was just panicking, but actually I don't think there is any reason to.
>
> (Emma, June 2015)

But by then, Emma's chances of achieving academic success that year had been gravely reduced.

Most participants identified listening as their main language-related problem. Reading was perceived as difficult and time-consuming, but not unsurpassable. All participants reported spending at least twice as much time studying in Dutch as they would studying the same matter in their L1. At least two students had translated extensive sections of Dutch coursework (Oksana) or entire syllabi (Stella) into their L1. Other participants (e.g., Gabriela, Océane, Clare) studied English or French course books that were similar to the compulsory Dutch ones.

As the year progressed, most participants felt that their receptive language skills were becoming better, but many perceived their productive language skills as stable or deteriorating. One of the reasons for this was a lack of writing tasks

during the academic year, and an absence of opportunities for speaking. At least four participants reported avoiding situations that would require them to speak Dutch (Elif, Leila, Hoang, Emma). These participants, and others (e.g., Guadalupe) considered the interviews a rare opportunity to speak Dutch freely.

The use of support systems

Last but not least, support systems turned out to be desirable but not in place. Throughout the year, the use of and experience with two different kinds of support systems was investigated: the university policy towards international L2 students, and the use of study support service. On the university level, there does not appear to be a clearly communicated policy for international L2 students. Participants typically felt that they did not really belong to the group of L1 students, or to the group of Erasmus students, but that there was no clear policy for the group of international students they associated with.

> We are an invisible group. People see Erasmus students, Flemish students, but international students are not very visible [...] If you want to get integrated, the label of international student is not what you need.
>
> (Alexandra, March 2015)

The support systems for international students during classes or examinations did not appear to be regulated in a systematic way, but seemed to vary from one professor to the next, and seemed to depend on the assertiveness of L2 students. Some participants who asked to use a dictionary, or to take an exam in English were allowed to do so, while others who had not asked were not able to take advantage of this opportunity. Similarly, some professors would allow some L2 students accommodations during examinations such as extra time or the use of English, while their colleagues would not.

> Some professors let us use a dictionary, while others do not. Some let us use the French version of the same law, while others do not. And some give a little more time during the exam.
>
> (Marie, March 2015)

> I asked [the philosophy professor] if I could do the exam in English and she said yes, but last semester I asked the same to a literature professor and she said no [...] And when I asked the student affairs department they said no. So I don't know who is right.
>
> (Gabriela, March 2015)

International students can use the study advisory services that are open to all students. Three participants knew of the existence of this service, regularly used it, and valued the support they received there. Others did not know it existed, or felt that the study support did not deal with the specific problems they encountered as international students.

Participants who left university, and their reasons to do so

The participants who left university voluntarily did so for a combination of interpersonal and institutional factors. Océane and Clara left primarily out of unhappiness with institutional factors, but stated that they had hoped to get to know more Flemish students. Both had underestimated the workload and had overestimated their ability to read course material in Dutch at the same pace as they did in French. For Océane, her inability to understand some of the lecturers, contributed to her decision.

> It's got nothing to do with Dutch. It's just too much work. [...] Yesterday I read 45 pages and it took almost thirteen hours of work [...] I'm just so tired. People expect me to work like this every day, but I'm sorry, I can't.
>
> (Clara, November 2014)

> In French, even when I don't listen carefully, I understand what is being said, but here when I don't listen carefully, I don't understand. I's much more exhausting [...] I wanted to come to Flanders, but I didn't realize that it would involve this much work, and that's the main problem for me.
>
> (Océane, November 2014)

Hoang left university during the June examinations for a number of reasons, but language and the feeling of isolation were the most important ones. For both him and Emma, the shock of being unprepared for the level of Dutch used in class caused a state of panic. Hoang perceived the situation as hopeless, found out that he could study in Germany the next year, and gave up before the end of the June examinations. In the second semester Emma learned to manage her language-induced panic, started to study routinely but did not catch up. In June she decided to leave university and pursue higher education at a Flemish university college.

Anastasia's case clearly showed that problems stemming from institutional factors can be aggravated by issues on the interpersonal level. She found the combination of not being able to adjust to the didactic culture, and of not connecting with her classmates too difficult. After she became ill in January her classmates sent her messages to complain about her not handing in an

assignment on time without inquiring about her health. At this point she decided to drop out:

> I don't like the university. The weird way of teaching and of dealing with students. I don't like that. [...] And my personal situation was the final drop. [...] I didn't want to quit. I think it's horrible, because the plan was to stay here as a student. But now I'm not studying anymore, but I can't work either because of the conditions in my visa [...] After the first month it was clear that I would not be happy here. I went to class and nobody spoke to me. When I asked a question, people replied politely, but that was it. And when I went to the student support service they didn't offer any help really.
>
> (Anastasia, February 2015)

Anastasia referred to visa requirements as a complicating factor in her decision to stop studying. Two students, Yazdan and Stella, involuntarily left university because of immigration issues. Yazdan needed a job because his allowance was reduced. The combination of working and studying proved too much in November 2014, so he decided to save money and return to university the following year. Stella, who was quite enthusiastic about her professors from the first interview on and was the only participant to have passed every exam, replied to an invitation for the March interview by saying that her visa had been revoked. At first she assumed that she had an agreement with the university to sit the exams in August and September (Stella, e-mail, March 25 2015), but later it appeared that the university required proof of permanent residency, which she did not have (Stella, e-mail, May 28 2015). By September it was clear that Stella would not return to Belgium:

> Unfortunately, I am forgetting Dutch more and more every day. [...] Nearly all I have done in the past three years has gone to waste. The only good news is that [...] I am now a qualified auditor in Armenia.
>
> (Stella, e-mail extract, September 17 2015)

DISCUSSION

Unavoidably, the results of longitudinal qualitative action research contain a myriad of uncontrollable variables. Often, the data stemming from such research can appear rather disordered, since the complexity of life is not easily captured in an abstract model (Leung, Harris, & Rampton, 2004). Nevertheless, in spite of all the important peculiarities and idiosyncrasies of every individual participant included in this study, some clear trends emerge. Utilizing the broad framework

proposed by the Douglas Fir Group (2016) in the analysis of the substantive amount of data gathered for this study, has facilitated the identification and organization of patterns that help to the explain the limited language gains as measured by the STRT test. In this discussion section, the research results will be interpreted in the light of previous findings, and the three levels of the Douglas Fir framework.

After eight months at university, the only significant performance difference between the test and the retest was a decrease in the amount of words used in the oral presentation task (p = .03, r = -.5). The amount of words used in the writing task had increased, however, with a difference that approached significance and a medium effect size (p = .07, r = -.4). This outcome reminds of findings by Knoch et al. (2015, 2014), who reported that the only significant difference in writing performance after one year and three years at an English-medium university was an increased written fluency, which was measured by number of words used. The current study adds further support to arguments challenging the assumption that exposure to L2 input will yield productive language gains (Ellis, 2003; Ortega, 2008). The data presented in this study further challenge the belief that international L2 students will make productive language gains by virtue of attending class in which the L2 is the medium of instruction (Knoch et al., 2015; Ranta & Meckelborg, 2013; Storch, 2009).

Educators in Flanders often use the metaphor of the "language bath", to support the belief that when L2 students are submerged in a context where the target language is the main or only language used, they will make language gains (Departement Onderwijs en Vorming, 2016). Previous authors have contested this metaphor, stating that L2 learners may drown in language baths (Van Avermaet & Slembrouck, 2014) – a statement which is supported by the longitudinal findings of the current study. Additionally, further eroding the validity of the language bath metaphor, this study shows that simply sitting in a language bath is not effective. Instead, learners need to get ample opportunities to meaningfully interact with their L1 peers (Elder & O'Loughlin, 2003; Serrano et al., 2012).

The results of the longitudinal interview data show, however, that most participants in the current study experienced problems on the interpersonal and on the institutional level, resulting in limited opportunities for meaningful interaction.

On the interpersonal level, it is clear from the results of the current study that interaction did not occur swiftly, smoothly or frequently, even if L1 and L2 students attended the same classes, or were involved in the same group assignments (see Ranta & Meckelborg, 2013). Issues of power and legitimate peripheral participation obstructed L2 students' access to the community of Flemish L1 students, which was perceived as impenetrable. Interaction is dialogic by definition however, and both parties share a responsibility for its success

(Kinginger, 2008). It is likely that many L1 students who were perceived as closed, did not have the intention to exclude their international peers, but were perceived as such by the $L2_F$ participants, who experienced a power imbalance. Indeed, situations that may appear rather innocent to members of the majority group could disproportionally impact members of minority groups (Walton & Cohen, 2007). Elena, for example, related an instance in which the classroom laughed and inferred that they had done so because she was from the Ukraine and must have therefore been stupid. Quite likely, her L1 classmates would have perceived that same situation differently, but for Elena this was the hard reality.

Quite a few participants faced similar issues, resulting in a renegotiation of their academic identity (see work by De Costa, 2011). Others, like Leila, Emma and Hoang, responded to a perceived language-related inferiority by withdrawing (see Duff, 2002). Many respondents at some point mentioned being impacted by (perceived) disapproving attitudes about them within the L1 community (Taylor, 1992). Some reacted to this by further distancing themselves, categorizing their L1 peers as young, childish, or spoiled. Similar dynamics have been reported by Norton (2013) and Pellegrino Aveni (2005), who observed that L2 learners may self-exclude from an interaction because of perceived or real power imbalances. Thus, even though the L1-L2 interactions were marked power asymmetries, the interview data show that it would be wrong to characterize L2 learners are passive objects who wait for their L1 peers to let them into their community of practice. Quite a few respondents made attempts to create a social network involving L1 students.

The second level, the one of institutional policies, can both facilitate and obstruct contact between L1 users and L2 learners (Holmes, Marra, & Vine, 2011; Douglas Fir Group, 2016). This study showed that the context of Flemish universities does not appear to be especially conducive to facilitating frequent L1/L2 interaction, or to promoting an empowered L2 identity. Especially during the first semester, participants did not experience institutional support, and they felt out of place in large classrooms with minimal interaction.

The participants included in this study felt that the university had a distinct policy for Erasmus students and for Flemish students, but not for them. The reason for this is clear: in policy terms they are not a distinct group at all. The current study did not find any examples of systematic support systems that cater to the needs of international L2 students. Individual professors did sometimes provide accommodations for L2 students' needs, but there was no clear system in place.

The interview data support the hypothesis (Douglas Fir Group, 2016) that the interpersonal and institutional levels are interconnected. The students who dropped out voluntarily, did so because of a combination of reasons related to the micro and meso-level: Many participants did not have a social network to fall

back on or a student network to help make sense of the rules and expectations at university, did not consider the professors approachable, and lacked the language to fully understand everything that was being said. Likewise, students who dropped out involuntarily did not have access to a powerful social network or to institutional support, to help them appeal the decisions that forced them to leave university, or the country.

In the course of this research no direct evidence of ideological forces – the macro level in the Douglas Fir framework – was collected. It is possible, however, to extrapolate a number of insights concerning the ideological structures present at Flemish universities from the institutional and interpersonal results. The predominant language ideology in Flanders has been described as territorial monolingualism (Blommaert, 2011; Blommaert & Van Avermaet, 2008; Van Splunder, 2015). Van Splunder (2015) writes that the language norm in Flanders is not simply Dutch however, but native-like Dutch, be it in dialect or in the standardized variety. He asserts that there is little tolerance towards language learners who do not attain that norm. Possibly, the fear of ridicule that withheld many participants from speaking in class (see also Duff, 2002), partaking in meetings, or interacting with L1 peers could stem from not being able to live up to this implicit norm. Additionally, national regulations that govern language use at university, limit the use of languages other than Dutch in order to reaffirm the importance of Dutch as a medium of instruction at university. Even though maintaining Dutch as an academically viable language is a valuable goal, strictly regulating the use of a language may foster a Dutch-only attitude among professors and students, which may reinforce already existing power asymmetries between L1 and L2 students. Similar dynamics have been observed in Flemish primary (Jordens, 2016; Strobbe, 2016) and secondary schools (Agirdag, 2010).

By the second semester, life at university had begun to improve, linguistically, socially and academically, adding support to the idea that international L2 students need time to adjust (Chirkov, Vansteenkiste, Tao, & Lynch, 2007; Kinginger, 2004; Ortega, 2008). Every international student included in this study needed time to negotiate the differences encountered in the new setting (Kinginger, 2010), to discover that their identity had been altered (Kinginger, 2004; Swain & Deters, 2007), and to gain acceptance into a new community of practice that they perceived as closed or unwilling to accept them. These results have a number of policy implications for Flemish universities.

First, this study shows that the sudden transition to university, which may be intimidating for any eighteen-year old, can be a daunting experience for an international L2 student. Consequently, in line with previous research (e.g., Kinginger, 2004), it took most participants a few months to get used to the new situation. Considering the difficulties they faced, it is quite striking that eight out

of sixteen participants (i.e., excluding Jessica and Chloé who left the study in November, and Stella and Yazdan who gave up because of migration issues) passed at least half of the courses they had registered for. Proportionally, that is comparable to the average study success in Flanders (Glorieux et al., 2015). It is not inconceivable that participants in this study would have fared better if they had experienced a smoother transition to university. Currently, Ghent University is piloting a program in which international students meet regularly during their first semester at university. During this time they also receive needs-based language support. Projects like this are vital for the increasing international students' chances of success.

Additionally, universities need to develop a clear and transparent support system. It should be clear for international students from day one which accommodations are available to them. Being able to take an examination in English or being granted more time to finish an examination should not depend on the assertiveness of a student or the personality of a professor.

Thirdly, this study reaffirmed that international students may not be immediately ready for the linguistic demands of university (Field, 2011). It has also confirmed that L2 students will not necessarily make language gains by virtue of attending Dutch-medium classes. With Byrnes, Maxim & Norris (2010), I would argue that it is the responsibility of the university to have clear L2 attainment targets in addition to entry requirements. If a university allows international L2 students when they have insufficient language to understand lectures, that university has a responsibility to provide opportunities for international students to develop their L2. Offering the language of instruction as a curricular course for L2 students is one way of doing so.

Lastly, international students who enter university may have gained relevant insights or expertise that could be put to good use in their own or in other programs. Leila's personal experiences in Haiti were certainly relevant to her political sciences program. Elena's background as an airplane engineer could have been turned into a useful contribution to class, and Oksana – L1 speaker of Russian – would certainly have been able to contribute to class in her Russian/English program of translation studies. Valorizing the potential of international students could positively impact the quality of the lessons, and would definitely make international students feel more at home in class.

EPILOGUE

Two years after conducting the first interviews the participants received the first version of this chapter (eight respondents replied: Leila, Alexandra, Gabriela, Alireza, Marie, Guadalupe, Oksana, and Elena). Their responses indicated that they had moved on, and that life at university had become somewhat easier after that first year. Oksana was about to graduate, and Gabriela, and Guadalupe were progressing academically:

> Today I am starting in the third year and after reading the paper I am extra motivated. For your information: I passed statistics 1 and 2 and I'm taking on a full study program, like a normal student. ☺
>
> (Guadalupe, text message, September 2017)

Others were excited about new professional opportunities. Leila had become a legal policy advisor at a government agency, and Alexandra – having found a job as an engineer – had decided to stay in Belgium for at least a few more years:

> I read your paper and it made me cry a little. I had already forgotten just how hard it had been for us. Now, for the first time I have been able to read about the experiences of other international students, and I was relieved to see that I wasn't crazy or alone.
>
> (Alexandra, mail, October 2017)

The future belongs not so much to the pure thinkers who are content – at best – with optimistic or pessimistic slogans; it is a province, rather, for reflective practitioners who are ready to act on their ideals. Warm hearts allied with cool heads seek a middle way between the extremes of abstract theory and personal impulse.

Toulmin, 2001, p. 214

PART 4
POLICY, CONCLUSION & IMPLICATIONS

> The first two parts of this dissertation were concerned with empirically investigating assumptions that support the university entrance policy for international L2 students. Until now, the focus was on proficiency levels, on representativeness, and on test equivalence. In the third part, which consists of one chapter, the attention shifts to what happens after the entrance test: do L2 students make gains in terms of academic Dutch language development, and if yes or no, why?

The final part of this dissertation includes three sections. Chapter 7 provides a discussion of the processes and mechanisms that have led to the Flemish university admission policy. Chapter 8 summarizes the research findings (Chapters 1 – 6), and Chapter 9 combines the outcomes of the empirical research with the perspective of policy makers in order to formulate realistic implications and recommendations.

CHAPTER 7
THE POLICY-MAKING PROCESS

The purpose of this chapter is to bridge the gap that sometimes exists between research and practice. It links the original research assumptions to the policy makers' perceptions, and shows to what extent empirical data may or may not impact policy.

In theory, policy-making is a straightforward, linear process: a perceived problem becomes part of the policy agenda, after which policy measures are developed and implemented. After some time, an evaluation of these measures leads to an adjustment, alteration or continuation of the existing policy (Howlett & Giest, 2013; Wilson, 2006). Laswell's (1956) policy cycle, briefly summarized here, remains influential in policy analysis studies, primarily because it conceptualizes a complicated process in a straightforward way. Like most models, however, Laswell's is an idealization (Jann & Wegrich, 2007). In reality, policy-making is cyclical nor linear, logical nor rational. It is influenced by a multitude of variables, such as budget restrictions, partisan tensions, or legal constraints (Van den Bosch & Cantillon, 2006).

Since policy is not easily captured in standard models, it is important to understand why and how policy is made in a specific context before relevant recommendations or implications can be formulated (O'Toole, 2000; Ross, 2008). Consequently, the purpose of this section is to trace the mechanisms that impact the Flemish university entrance policy. This chapter relies on interviews with policy makers, conducted after the empirical data presented in this dissertation had been analyzed. It addresses a gap in the language testing literature by showing how and why university entrance language requirements are determined.

EXAMINING UNIVERSITY ADMISSION POLICIES

In the context of higher education, few studies have examined the real-world conditions under which admission policies are shaped. What is more, studies in this field have traditionally adopted a somewhat positivistic technocratic model, which focuses on quantifying the extent to which a policy reaches its objective, without necessarily considering real-world limitations (Fischer, 2007; Howlett & Giest, 2013; Vedung, 2013). Typically, these studies examine the effectiveness and side effects of a policy. In an important publication in this tradition, Wainer

(2011) focused on the use of SAT results in the university admission policy of North American universities. The quantitative analyses uncovered fundamental flaws and inconsistencies in the assumptions that support university admission policies. Wainer did not focus on language tests, but his conclusions were a warning call about the dangers of conducting policy on the basis of misguided or untested assumptions. Similar publications in the field of educational policy often find evidence that partly or completely disproves the very premise on which a policy relies (e.g., see Ball, 2015; Borg, 2006 for a discussion).

There are only a handful of studies that focus specifically on the language requirements in university admission policies. The assumptions behind a university entrance policy for international L2 students have not often been the topic of extensive language testing research (McNamara & Ryan, 2011). Studies that do touch upon this field mostly focus on score use. O'Loughlin (2011, 2013) showed that university admission officers are not always aware of the meaning and scope of a test score, and primarily desire clear-cut, straightforward information. Green (Forthcoming) demonstrated that the information provided by language test developers about a test's CEFR level was not as clear-cut as it seems. He concluded that equivalence between English university admission L2 tests that share the same CEFR level cannot be assumed, since the procedures used to link to the CEFR may differ substantially. Chapters 3 and 4 of this dissertation offer further backing to Green's point.

Importantly, research that examines university admission language requirements or tests typically adopts an empirically-driven research paradigm, rather than the pragmatic perspective of policy makers (Howlett & Giest, 2013; Jann & Fischer, 2007; Wollmann, 2007). A clear example of this can be found in the language assessment literacy literature. Fueled by evidence of intended or unintended misuse of language tests and language test scores (Fulcher, 2012a; O'Loughlin, 2011, 2013; Spolsky, 2008; Taylor, 2009, 2013), this discipline is concerned with the assessment-related competences required by various stakeholders in the language testing process. Importantly, however, authors in this field tend to presume that the language tester is at the heart of the process (Malone, 2013; Taylor, 2013). Consequently, the approach taken in the language assessment literacy literature is explicitly top-down: "those who need to develop such literacy are likely to have less time and energy to spend seeking out what is relevant and useful to their requirements; the onus of responsibility for making key information more accessible must surely lie with those who already know where it is located" (Taylor, 2013, p. 408).

Recent policy evaluation research strongly suggests that academic policy evaluation initiatives are not likely to bring about real-world change (Wilson, 2006). To have impact, researchers should be aware of the exact context in which a policy is set, be attentive to the real-world constraints (Ross, 2008), and accept that policy-making is messy by default (Ball, 2015). If they are not, policy advice

might be too disconnected from reality to have any impact at all (Bovens, 't Hart, & Kuipers, 2006).

RESEARCH QUESTIONS

Before formulating policy recommendations (Chapter 9), this section maps the constraints and conditions that impact the university admission policy in Flanders. By investigating two explorative research questions, it examines why the Flemish language-related university admission requirements are what they are:

RQ1 *How is the university admission policy for international L2 students at Flemish universities made?*

RQ2 *What are the commonly held assumptions behind this admission policy?*

PARTICIPANTS & METHODOLOGY

Participants

To select participants, a purposeful sampling strategy was used. At the end of the data analysis, in November 2016, the vice-deans and educational directors[1] of the five Flemish universities were asked to identify the members of staff who are directly responsible for the admission criteria regarding international students at their university. All participants were senior members of staff, directly responsible for the admission policy at their institution.

Table 8.1. Policy maker respondent codes

University of Leuven	UL1	UL2		
Ghent University	UG1	UG2		
University of Antwerp	UA1	UA2		
University of Brussels	UB1	UB2		
University of Hasselt★	UH1	UH2	UH3	UH4
Flemish Government	FG1	FG2	FG3	

Note. (★) Four more members of staff were present, but did not partake in the interview.

[1] Unlike the other Flemish universities, Ghent University does not have a vice-dean for education. Instead, the official title is Director of Educational Policy.

Importantly, since the policies at the Flemish universities are partly determined by a Flemish decree (Vlaamse Regering, 2013), three senior policy makers at the government level were also recruited. These respondents were specifically responsible for guidelines concerning university admission issued by the Flemish government. All policy makers at both levels agreed to participate, and as such the research population (N = 15) represents the full real-world population. In order to guarantee anonymity, all respondents will be referred to using a code, consisting of the acronym for their affiliation, and a number (see Table 8.1).

Data collection & analysis

All interviews except for one, which was limited to 45 minutes because of the respondents' schedule, took more than one hour (Min: 49 minutes, Max: 91 minutes, Median: 74.5 minutes). Every interview was structured, and built around three central components. First, the respondents were asked to explain how the university entrance policy regarding international students is shaped. Next, every university entrance requirement in place at a given institution was discussed. In this phase of the interview the assumptions that drove this research were checked with policy makers. Assumption 2 (test representativeness) was not included in the interview scenario, because this was not considered a claim made by score users, but by test developers. Respondents were prompted to explain the nature, purpose, and perceived effectiveness of each requirement (see Table 1.2). In the last component of the interview, the main research conclusions of this research were discussed with the respondents. Every interview was audio recorded and transcribed. The transcriptions were analyzed using an a priori coding tree consisting of three main branches (see Table 8.2).

Table 8.2. Data coding categories

Branch	Topic	Code
1	Policy-making process	impacting variables, empirical foundation
2	Policy enactment	goal, effectiveness, post-admittance policy
3	Policy assumptions	B2 level requirement, STRT-ITNA equivalence, Flemish students' language proficiency, post-entry language gains, 60 credits and language proficiency

The branches of the coding tree correspond to the interview scenario, and primarily concern factual data. The first two branches focus on how policy is made and enacted (relying on policy evaluation literature, e.g., Jann & Wegrich, 2007) and the third branch investigates to what extent Assumptions 1, 3, 4, and 5 are upheld by the policy makers.

RESULTS

How policy is made at government level

The admission requirements for international L2 students at all Flemish universities share three guidelines, which originate from a Flemish government decree known as "the codex" (Vlaamse Regering, 2013). This document briefly stipulates that universities *may* use the following documents as sufficient evidence for the admission of international L2 students: (1) a language test result, (2) proof of having successfully completed one year in a Dutch-medium secondary school, and (3) proof of having achieved 60 credits in a Dutch-medium higher education program. The codex does not specify a required language level, and neither does it identify which tests are accepted. A more recent decree, which concerns international L2 professors is noticeably more specific, but the lack of specification in the codex is not necessarily the consequence of a deliberate strategy.

FG2 I think it's primarily a difference in timing. [The codex] is very old. Actually, it's always been like this. And now, with the new regulations for professors they went much further.

Because the codex has been in use for quite some time, the respondents did not know what it is based on, or where the three requirements came from: "We have been working here for a pretty long time, but this rule has been around longer [...] To us the rules always seemed logical. Actually, they have never been questioned in all those years" (FG1). Policy measures are not routinely evaluated, but reconsidered when universities or political stakeholders raise concerns, and the admission requirements have through the years "simply been reused" (FG2). When policy texts are revised, the impact of empirical research is minimal, compared to the impact of stakeholders.

FG3 Research is often used a little selectively, like when people want something to become policy.
FG1 If we wanted to change policy we wouldn't first do or order a study.
FG2 What could happen is that policy advice is based on a scientific study [...] But there are always political negotiations, and all stakeholders are involved.

How policy is made at university level

All respondents defined the goal of the university admission policy in a very similar way: To select students who have a sufficient level of language proficiency

to be able to attend a Dutch-medium university program. When asked about the measures in place to pursue that policy goal, all respondents referred to the codex as the key source for the requirements. Since universities are under no legal obligation to follow the codex, however, most institutions have identified exemptions for students at the master or postgraduate level (see Table 8.3). Only the University of Brussels has exemptions for bachelor students.

Respondents at two universities mentioned relying on trends in pass and fail rates when making policy decisions, but not when it came to the language requirements. No respondent referred to empirical research as a reason for adjusting the language requirements for specific programs, but not because of fundamental objections.

Table 8.3. Exemptions from admission requirements at Flemish universities

University of Leuven	• Faculty or program can drop language requirements for master students
Ghent University	• Faculty or program can drop language requirements for master students
University of Antwerp	• Faculty or program can drop language requirements for postgraduate students • Faculty or program can exempt students with partial study load (so-called *credit contracts*)
University of Brussels	• All students from a Belgian French-medium secondary school • Program directors can decide to allow individual student after a review of their file
University of Hasselt	/

UL2 Most likely we will never be able to implement very big changes but we are able to make recommendations supported by research. It will still remain to be seen if the proposal is passed though […] in the end it's a game of politics.

To a large extent, policy decisions at university level appear influenced by both internal stakeholders and external factors. The exemptions listed in Table 8.3 are the result of internal pressure from specific faculties or programs. University admission officers cannot design a policy without taking into account the views and aims of powerful internal stakeholders:

UA1 We have had a lot of debate about [an exemption for certain students] with professors who wanted to attract specific profiles.

UA2 That's right. Is everybody happy with these exemptions? I have to be honest: no. But it has been decided that they want to continue to allow for this exemption.

Quite a few respondents referred to admission requirements that were purposefully vague. "I think there are supposed to be small flaws in the system", UL2 stated, before referring to instances of professors admitting students who had not passed an accredited test: "I try to stand my ground then [...] and sometimes I can, sometimes I can't".

External events that are beyond the control of a university may also prompt policy changes. For example, after 2012 the University of Brussels had dropped all language requirements for international L2 students. In 2015 the decision to reinstate them was taken because the university was suddenly faced with an influx of German students. From one year to the next, more than half of the freshmen in psychology and biomedical sciences were German students: "These students were coming to our university because of a reform in their secondary education system. [...] These people were looking for solutions, and they came here. And then we needed to fix that situation" (UB1). Another external influence is the policy adopted by other universities. The University of Hasselt lowered the required entrance level to match the policy of a larger university nearby: "it was because Leuven had B2 as well back then. And also because we thought C1 was rather high, and then we checked the CEFR and thought B2 would be sufficient" (UH2).

Commonly held assumptions

At every university the default entrance level is B2, but the reasons given for using that level do not necessarily refer to empirical research. In spite of the absence of empirical backing, all respondents considered B2 an adequate threshold level to differentiate between students whose language proficiency is likely to be an obstacle to academic success, and students whose proficiency will not prove to be a hindrance. At the same time, respondents pointed out that there are many other variables influencing a student's academic success: "It's not just language. It's a very complicated process. I think everybody thinks 'we have to do something, so let's do this'." (UH3).

At all universities, both STRT and ITNA are legally equivalent. Respondents at three universities also assumed level equivalence between the two tests, but without a clear rationale: "*we consider them equivalent. I don't know why. Probably because both are at the B2 level according to something or somebody*" (UL2). Respondents at other institutions noted that for them, the legal perspective was the only one that mattered.

UG1 We don't actually wonder about equivalence [...] What matters for us is that we have all the documents we need for registration [...] We assume that both tests are minimally B2.

UG2 Legally they are equivalent, and that is our approach.

Most respondents believed that students graduating from Dutch-medium secondary education are at the B2 level. At the University of Brussels the participants were not so sure, because of the predominantly French population in Brussels. Few respondents felt sure that the requirement of having successfully attended one year at a Dutch-medium secondary school would be reliable proof of B2 proficiency. At the same time, this requirement has not been altered in any university admission policy.

UA1 It's always been in there [...] You could doubt this requirement, probably [...] I have always wondered if we could make it more strict but I know there would be a lot of internal resistance if we would do that.

UA2 No! We should accept the guidelines in the government decree, and not make them stricter. Let's give [students] a chance if we can. Their language proficiency may improve at university.

As UA2's last comment indicates, respondents generally expected international L2 students to make language gains by virtue of attending Dutch-medium classes. Most universities offer academic language classes for students who experience language-related problems, but since there is no general post-admittance policy for international L2 students, these classes are mostly geared towards the general (i.e., L1) student population.

When asked to assess the effectiveness of the policy measures to reach the policy goal, all respondents were hesitant, since they generally lack the means to measure effectiveness in a precise way. At most institutions, there were no clear statistics of international L2 students who study in Dutch. No institution had a post-admission policy for these students, so it is largely impossible to track them. If universities focus their attention on international students, they tend to concentrate on those who attend English-medium programs. Orientation days for international students, for example, do not normally focus on those students who will study in Dutch: "*I think the group of students is too small and hard to reach and that's why it doesn't happen* (UL2)".

Finally, during the third part of the interview, the research results were discussed. This component of the interview generated a lot of interest, and after every interview participants stated that they would take the data into consideration. Participants at every university sent follow-up e-mails, expressing interest in the research, asking to be kept informed. Nevertheless, it was clear that this was no guarantee for the findings to find their way to policy: "*it all*

depends on what is politically achievable at a given moment. […] What happens next is a whole process of negotiations and talks. And in the end, the original proposal may look entirely differently" (FG2).

DISCUSSION

The interview data consistently show that the Flemish university admission policy, like any real-world policy, does not rigidly follow a linear or cyclical logic, and it does not rely on routine evaluation (Jann & Wegrich, 2007). Instead, policies are adjusted when problems need to be solved, or when important stakeholders want change.

This study confirmed that the university admission policy is the result of a series of pragmatic rather than empirical or logical decisions (Ball, 2015). Universities follow the codex, except when they do not. They may sometimes use empirical data, but at other times may not. Tellingly, no respondent gave empirically founded arguments to argue why specific language requirements had been adjusted. In reality, these requirements were used to flexibly control student access to certain programs (see also Chapter 1). At the University of Brussels, the decision to first dismiss and later reinstate language requirements was driven entirely by practical considerations regarding the student population. Similarly, the University of Hasselt lowered the requirement from C1 so they would not lose students to Leuven. At the University of Antwerp, certain programs have no language requirements, simply because professors do not want to lose certain students because of them.

If policy is essentially about the use of power, as Wilson (2006) argues, then policy making is about making compromises that take into account the diverging interests of different parties. The same dynamic is demonstrably present at Flemish universities, where program directors are important drivers of the Flemish university admission policy. This may explain why some admission requirements for international L2 students have remained unchanged and unchallenged for years: International L2 students studying in Dutch are too small a group to be powerful, and too dispersed to be noticed by a stakeholder.

The central assumptions in this dissertation were mostly confirmed. Among the respondents there was a consensus was that B2 is a satisfactory minimum level for university admission. Secondly, while policy makers at two universities considered STRT and ITNA linguistically equivalent, there was unanimous agreement among all respondents that both tests are legally equivalent. The issues of level equivalence and construct equivalence are not necessarily equally relevant to policy makers, since their frame of reference appears to be primarily oriented towards legality. Thirdly, there was general consensus that students

with a Flemish high school degree can confidently be assumed to have B2 language proficiency. Lastly, even though this is not a language requirement, quite a few respondents remarked that they would expect international L2 students to make language gains by virtue of studying at a Flemish university.

CONCLUSION: A DIFFERENT PARADIGM

No respondent believed that the admission system at their university was watertight. Nevertheless, it was felt that the admission criteria served their purpose, because they were an acceptable compromise. The interview data largely confirm that policy comes down to fixing problems (Ball, 2015), and in this patchwork of pragmatic compromises empirical research is of little importance.

One and the same problem looks different from different angles, and policy makers and researchers may come up with very different solutions to the same issue (Goodin, Rein, & Moran, 2006). Echoing this paradigm schism (Howlett & Giest, 2013; Jann & Fischer, 2007; Wollmann, 2007), this study showed Flemish university admission policy makers are pragmatists. Researchers, on the other hand may not always take into account real-world constraints or political strategies. Language assessment literacy authors, for example, may adhere to a paradigm that is quite distant from the one used by policy makers. Outlining assessment competency profiles that list the specific testing expertise that would be required of university admission officers (Taylor, 2013), does not appear to match the day-to-day reality of policy makers. Similarly, schooling admission officers in matters of validity, in order to ensure that they can make informed individual decisions (O'Loughlin, 2013) may adhere to a somewhat idealistic paradigm that is quite distant from the pragmatic one used by policy makers.

That does not mean that policy makers should not be informed about empirical results. The information should, however, meet their frame of reference and their day-to-day reality, rather than the researchers' ideals. Proposals from researchers that are premised on academically-oriented paradigms may too distant from reality to have an impact. Nevertheless, since policy makers are limited in what they can do by an invisible network of interests and power politics, empirical results may not have the impact that researchers may wish it to have, even if results are communicated in an appropriate way. It is safe to say that the responsibility for a university admission policy may reside with the policy maker, but the power to change it does not.

CHAPTER 8
SUMMARY & DISCUSSION OF THE RESEARCH FINDINGS

In Flanders, Belgium, international L2 students are typically required to prove B2 language proficiency in Dutch. Different universities accept different kinds of proof of B2 ability, but all institutions accept a language test certificate by STRT or ITNA. Additionally, international L2 students with any Flemish secondary school degree are allowed to enroll, and students who have already obtained sixty credits at a Dutch-medium Flemish university or university college can register at another university without having to prove B2 ability again. This research project investigated the Flemish university entrance policy by addressing five different research goals. The first four goals were based on four assumptions drawn from the entrance requirements that all Flemish universities share. The fifth considered language gains made by international L2 students after admission.

EXAMINE THE EMPIRICAL SUPPORT FOR THE B2 LEVEL AS AN ENTRANCE REQUIREMENT

Summary of the findings

The first two chapters addressed the matter of using the B2 level as the university entrance threshold level. The first chapter reported on structured interviews conducted with 30 informed respondents from 28 European contexts that have (quasi) autonomy over educational matters. The findings show that throughout Europe, B2 is the most commonly used level to determine university entrance. In twenty of the 22 European contexts that use CEFR-related language requirements to determine university entrance, B2 plays an important role in the university entrance requirements for international L2 students. In ten of those contexts, it is the only requirement. In ten other, B2 is one of the requirements, but for some programs A2 (n = 1), B1 (n = 1), or C1 (n = 8) might be the minimum entrance level. At the same time, only three respondents out of 30 were assured that B2 users would be able to function linguistically at the start of university. Additionally, the required entrance level was based on empirical research in only one out of 23 contexts that use language tests to determine university entrance for international L2 students. On the whole, the results from this study strongly suggest that the B2 level has become the default university entrance level in Europe without much empirical backing.

The second chapter investigated whether the B2 level corresponds to the minimum language requirements at Flemish universities. The results show that the university staff (N = 24) considered the B2 level vastly insufficient for listening and reading, but acceptable for writing. Similarly, the international L2 students (N = 31) consulted for this study all struggled with the real-life listening demands of university. All respondents reported problems with understanding their first lectures, and a few had understood nothing at all – mainly because they were not prepared for the variation in accents and pronunciation styles encountered in real life. Actually, the C1 descriptors in the CEFR appear to match the real-life receptive requirements more than the B2 descriptors do. For reading, the C1 level states "lengthy, complex texts likely to be encountered in social, professional or academic life" (Council of Europe, 2001, p. 70). Similarly, the C1 descriptor for listening mentions unfamiliar accents, and following "extended speech even when it is not clearly structured and when relationships are only implied" (Council of Europe, 2001, p. 66).

The international L2 students reported fewer problems with reading and writing, primarily because these skills typically allow language learners to deal with input at their own pace. Typically, reading in Dutch was estimated to take twice as long as compared to reading in the L1.

Discussion

The data presented in this dissertation have reaffirmed that the CEFR has fundamentally altered university entrance language testing in Europe (Little, 2007). The wide uptake of the CEFR could be considered a good thing, and proponents have pointed out the benefits of using a common language to describe language proficiency levels (North, 2014a, 2014b, 2016). Without wishing to deny the positive impact of the CEFR on teaching and curriculum development (e.g., the focus on a can-do approach), it is important to remain aware of potential shortcomings in the way the CEFR is used in high-stakes testing. Throughout this dissertation, two important CEFR-related risks recurred: normative use of threshold levels, and reification of the B2 profile.

The first risk – normative use of levels – implies that the B2 level is required for university admission, simply because it is B2. It becomes the norm, not because empirical data show it to be adequate, but because it already is a norm in other contexts.

The CEFR itself has been amply criticized for the gaps in its empirical foundation (Alderson, 2007; Fulcher, 2012b), for not incorporating insights from empirical SLA research (Little, 2007), and for not relying sufficiently on actual learner data (Hulstijn, 2007). This criticism has been party acknowledged by the CEFR's authors, who estimated that some ten percent of the illustrative

descriptors (i.e., 23 C-level descriptors, the entire *orthographic control* scale, ten *sociolinguistic appropriateness* descriptors, and two *phonological control* descriptors) had not been empirically validated (North, 2014a). Currently, an initiative is underway to add empirical support to certain descriptors, while also developing new scales (e.g., *mediation*). These flaws in the empirical foundation of the CEFR itself have been the subject of much scrutiny, beginning rather soon after its publication (e.g., Fulcher, 2004). It would seem that a framework that is contested empirically would be used for high stakes purposes only after careful analysis. Yet, in the overwhelming majority of European contexts surveyed, this does not appear to happen. The B2 level appears to have achieved a special appeal in the context of university admission, without a clear scientific or empirical rationale to back its omnipresence.

Normative use of the B2 level can be observed in policies, but also in tests. Clear cases in point are the Flemish STRT and ITNA tests, which were developed with the B2 level in mind. To some extent the typical features of this level seem to have driven the operationalization of these tests. In its validity argument, ITNA justifies most of its item types by referring to what a B2 learner of Dutch can do (Interuniversitair Testing Consortium, 2015). In STRT the B2 level orientation was requested by the funding organization (Nederlandse Taalunie, 2013). Especially in the selection of input material it becomes clear that STRT may have relied more on B2 descriptors than on an analysis of target language use demands. Even though the STRT listening and reading prompts may not always reflect the challenges of real-life receptive demands, they match the B2 descriptors quite well indeed.

Apart from normative CEFR use there appears to be no clear motivation to create a university entrance test that tests every skill at the B2 level. Actually, one of the goals of the CEFR was to offer an alternative to the idea that one person had one kind of uniform proficiency level (Krumm, 2007). In a 2011 paper, Hulstijn argued that demanding an even CEFR profile in an academic context would most probably not correspond with real-life requirements. The research discussed in Chapter 2 offers empirical support to Hulstijn's assertion: in Flanders, B2 may be an appropriate threshold level for writing and possibly speaking, but for listening and reading a higher level – or one that incorporates more features of real-world language use – would be closer to the real-life demands.

The observations made in the paragraphs above tie in with the second risk: reification of the CEFR levels. This threat was coined by Fulcher in a 2004 paper, and remains true thirteen years later. A logical fallacy first termed "the fallacy of misplaced concreteness" (Whitehead, 1925), reification occurs when something abstract is treated as if it were something concrete. When the B2 level is treated as something that is an exactly measurable entity such as temperature

or weight (e.g., De Jong, 2013, 2014), it is being reified, and this may over time endanger the CEFR's credibility and usefulness (Hulstijn, 2015).

In this research no data regarding the linking procedure of STRT or ITNA was consulted. Both tests used analogous familiarization-specification-standardization procedures (Figueras, North, Takala, Verhelst, & Van Avermaet, 2005) well enough to pass an ALTE audit. Nevertheless, being linked to the same CEFR level far from guarantees equivalence, as is clear from Chapter 3 and 4. On the surface, using the same CEFR levels across tests may seem to increase score comparability and score transparency, but in practice the difficulty levels of two tests that share the same CEFR level may diverge substantially. The fact that two tests that are linked to the same CEFR level may lack equivalence should not surprise informed researchers: B2 covers quite a range of performances, and the CEFR is hardly precise enough to act as a ready-made measurement scale (Galaczi, ffrench, Hubbard, & Green, 2011; Harsch & Martin, 2012; Weir, 2005b).

This should not necessarily be a problem. The problem only arises when CEFR levels are somehow believed to be "true", or when stakeholders assume that tests are equally difficulty because they have the same CEFR level. Using tests as if they were equivalent, or implying that they are, only because they share the same CEFR level, does not qualify as responsible score use, and policy makers should be warned against it. Measuring true score equivalence entails conducting test-equivalence studies by means of equipercentile ranking (ETS, 2010), regression analysis (Zheng & De Jong, 2011), or Rasch (Fulcher, 1997). Another recommended approach, which does not necessarily reduce the need for equivalence studies, is to develop test specifications based on the actual needs and requirements of the target context, and to link these tests to the CEFR later.

Overall, the evidence suggests that the empirical basis for using the B2 level in the first place, is rather thin. It also appears that using the B2 level as an overall requirement for university entrance does not correspond to the real-life demands. In reality, language demands are often unevenly distributed across skills (Hulstijn, 2014). Based on these results, it is recommended to match the receptive test tasks to the real-world requirements. This could imply that that the level requirements for receptive skills move closer to C1. As it stands, international L2 students appear quite likely to enter university with receptive skills that do not prepare them sufficiently for the real-life demands in the target context.

COMPARE REAL-LIFE LANGUAGE REQUIREMENTS AT FLEMISH UNIVERSITIES TO STRT AND ITNA OPERATIONALIZATIONS

Summary of the findings

The second chapter did not only focus on the B2 level as an entrance requirement at Flemish universities – it also compared the operationalization of STRT and ITNA to the real-life demands of academia. Speaking, for example, is of comparatively minor importance at university in Flanders. The university staff ranked it as the least important skill, and most respondents spoke Dutch rather sparsely in their daily lives. Nevertheless, speaking tasks feature quite prominently in STRT and ITNA. At times, the content and requirements of the tests seem to rely more on general conceptualizations of what language for academic purposes comprises, than on a thorough analysis of the target context. Additionally, the study did not generate data to convincingly suggest that there is a clear link between academic success and scores on STRT and ITNA. The sample size used to draw that conclusion was rather limited, but the results are in line with other international predictive validity studies that show the weak relationship between test scores and academic performance (Cho & Bridgeman, 2012; Hill, Storch, & Lynch, 1999; Kerstjens & Nery, 2000; Lee & Greene, 2007). These previous studies solely considered the academic performance of successful language test candidates. In the current research, a small cohort of students who failed one of two language tests were tracked during their first year at university.

Discussion

Throughout this dissertation, different kinds of data were collected that can be used to investigate the validity argument of STRT and ITNA as university entrance language tests. It is important to reiterate that test developers are required to substantiate a validity claim in any context they explicitly promote or can reasonably foresee (Kane, 2013). In other words: the fact that STRT is also used as an entrance test to university colleges and that ITNA is also used as a course-final level test, does not diminish the need for test developers to provide a convincing validity argument in every single context.

In a forthcoming publication, Kane, Kane, & Clauser (2017) lay out a framework for validating credentialing tests. These tests are used to determine whether candidates possess the right knowledge, skills, and/or judgment to be allowed access to a target domain. The framework provides a useful perspective on validation, which can be used for the purpose of this dissertation.

Interpretation/Use Arguments (IUA) for credentialing tests, Kane et al. (2017) argue, typically rely on minimally four inferences: a *scoring inference* (which links a performance with a score), a *generalization inference* (which translates all item or task scores to an overall score, often using statistical modeling), an *extrapolation inference* (which infers real-life performance on the basis of a test performance), and a *decision inference* (which determines the performance level that marks acceptable competence). These inferences are linked to four suppositions, which are paraphrased and discussed below.

(1) The test tasks incorporate skills that are essential for the target context.

Table 9.1. STRT & ITNA task types, to important skills to master when entering university

	STRT						ITNA						
	Argumentative writing from audio	Summarize lecture	Argumentative writing from text	Summarize paper	Oral argumentation	Presentation	Language-in-use, vocabulary	Language-in-use, grammar	Reading comprehension	Listening comprehension	Listening comprehension, dictation	Oral argumentation	Presentation
Compose a logical argumentation	★		★		★								
Take class notes	★	★											
Express ideas accurately	★	★	★	★	★	★							★
Grammatical accuracy	★	★	★	★	★	★		★				★	★
Understand general academic lexis	★	★	★	★	★	★	★		★	★	★	★	★
Understand coherence & cohesion	★	★	★	★	★	★			★	★		★	★
Understand implicit message	★	★		★					★	★			★
Understand scientific text as a whole				★					★				
Look up information													
Summarize long text				★									
Summarize multiple sources													
Understand scientific text in detail				★									
Describe graphs & tables						★							★
Give a presentation						★							★

In Chapter 2, the university staff respondents and the $L2_F$ participants were asked to list the skills they considered essential for students to master at the start of university. Table 9.1 includes those skills in four categories. The first category shows the skills that were considered essential by both university staff and $L2_F$

respondents. Below the first dashed line are the skills that were considered essential by of one of both groups. The third category lists skills that were selected once or twice in each group, but were not considered important by either. The last category contains the skills that were never selected by any L2 or university staff respondent. Within each category, the skills are placed in an alphabetical and non-hierarchal order. If a given STRT or ITNA task type (Appendix 1 and 2) operationalizes a certain skill, or uses it in the rating criteria, it is marked with a "★". The table was constructed by relying on the task specifications, on input from the STRT and ITNA teams, and on the validation reports written by the test development team in preparation for the ALTE audit (CNaVT, 2014; Interuniversitair Testing Consortium, 2015).

It is important to stress that Table 9.1 is a rough outline, which primarily serves to display whether a test operationalizes in some way the LAP skills that were introduced in Chapter 2. It offers a binary picture, and does not convey shades of importance. For example, grammar and vocabulary are pivotal to ITNA's construct and largely account for its difficulty level, as was concluded in Chapter 3 and 4. Judging from the table, however, it may appear that STRT assigns greater importance to grammar and vocabulary, because they consistently used as a rating criterion. While Table 9.1 does not convey nuance, it does display quire clearly whether ITNA and STRT in some form consider academic language skills that are considered essential for the target context.

Two essential skills are not operationalized in ITNA's B2 test, in the sense that they do not impact a test-taker's score: "*compose a logical argumentation*", "*take class notes*". At the B2 level, ITNA does not include productive writing tasks, and the rating criteria in the oral component are focused on linguistic quality, rather than content.

In some way, the STRT tasks incorporate every essential skill, but this does not mean that every skill is operationalized perfectly. The "scientific texts", for example, are popularizing rather than academic, and the listening prompts differ substantially from actual lectures. In ITNA, the reading tasks also rely on popularizing sources, but the audio prompts are more in line with natural – but not necessarily academic – language use, since they are actual radio fragments. Because STRT pays considerable attention to content criteria, it consistently appeals to the essential skill "*express ideas accurately*". Nevertheless, the way in which this skill is operationalized is probably below the level of real-life academic language: Chapter 3 showed that STRT's written argumentative tasks (in which accurate expression of ideas is of pivotal importance) and its content criteria are disproportionally easy.

(2) The absence of certain test tasks would make a substantial difference in real-world practice effectiveness.

The primacy of speaking in STRT and ITNA contrasts to some degree with the moderate importance of speaking skills in the target context. The prominence of speaking tasks in the tests may have to be reconsidered, or the approach that is adopted (i.e., both tests could focus more on social interaction, or on discussing a topic chosen by the candidate). In any case, the current ITNA design, in which access to the oral component is granted based on the computer test score, seems to give an aura of importance to speaking, which does not correspond to reality.

It is also doubtful whether omitting the dictation task from ITNA would make a substantial difference in real-world practice effectiveness, since it misfits the Rasch model composed of all STRT and ITNA's written tasks (Infit MnSq 1.71 – see Table 4.5) and does not operationalize any of the essential LAP skills. Lastly, there are clear indications of redundancy in the written STRT component. The Rasch model does not reliably differentiate between the two summary tasks (Listening, Measure = -.01, SE = .08; Reading, Measure = -.04, SE = .08) and the two argumentation tasks (Listening, Measure = -.59, SE = .11; Reading, Measure = -.51, SE = .10), one of which overfits the Rasch model (Argumentative writing-from-reading Infit MnSq = .41). In a writing test that takes three hours, it might be beneficial to omit tasks that may be considered redundant.

(3) The scoring is accurate and consistent.

Before conducting analyses on the test scores (Chapter 3 and 4), the rating procedures of STRT and ITNA were checked. The Rasch rater analysis of the $L2_F$ performances on STRT showed only minimal differences in rater severity (.73 logits). At .35, the reliability with which the model could differentiate between raters' severity was low, which, together with the non-significant $X^2(4) = 3.5$ (p = .47), showed that the assumption of rater equivalence could be upheld. There were no direct indications that cast doubt on the accuracy or consistency of STRT ratings.

A python script was used to determine which answers ITNA's automated scoring tool considered correct or incorrect, and to verify the scoring algorithm of ITNA's computer component. Since only one, rather minor, inconsistency was found, it was decided that the scoring procedure of ITNA's computer section was satisfactory. The accuracy and consistency of the scoring procedure used in ITNA's oral component is hard to assess, since two raters reach one jointly agreed-upon score. ITNA's internal audit report (Interuniversitair Testing Consortium, 2015) only includes information regarding the correlation (r = .90) inter-rater agreement (k = .73) between two raters. The report also mentions

research regarding the consistency across test centers. A comparison of the total scores assigned in the different ITNA test centers yield non-significant results on the Kruskal-Wallis test (p = .39). The report takes this as evidence of rating consistency, but it is possible that other variables (e.g., stronger candidates would mask stricter raters) have influenced this outcome. Consequently, it is impossible to make any reliable claims regarding the accuracy of ITNA's oral scoring procedure.

Chapter 4 offered additional data regarding the consistency of the scoring procedures across tests and showed that the same criteria were operationalized very differently in STRT and ITNA. Such a comparison does not yield information regarding internal scoring consistency, but it does underscore the uncertain nature of determining what "appropriate" scoring may entail.

(4) The passing score is appropriate.

If an entrance test is considered a measure of *necessary* (i.e., what is minimally required) rather than *effective* performance (i.e., what should ideally be mastered), candidates who struggle with test tasks can be expected to experience problems in the target context. On the other hand, not all candidates who pass will be expected to thrive (Kane et al., 2017). This line of reasoning appears generally true for STRT and ITNA, though the sample size used to assess it in this research was too small to make any definitive claims.

Table 9.2. STRT & ITNA result vs. academic success

		Academic success	
		<50%	>50%
STRT	Fail	1	0
	Pass	7	8
ITNA	Fail	2	2
	Pass	6	6

Corroborating previous research (e.g., Cho & Bridgeman, 2012; Lee & Greene, 2007), no significant relationship was found between language test scores and academic success. The effect sizes for both tests were comparable (STRT, r = -.115; ITNA, r = -.120). A crosstab (see Table 9.2) shows the relationship between STRT and ITNA pass/fail outcomes and academic success (defined by having passed more than 50% of the credits). Based on the sixteen respondents who did not leave the project prematurely because of visa issues or without stating a reason, ITNA assigned two false negatives and six false positives, whereas STRT assigned zero and seven respectively. If we include the $L2_F$ respondent Stella, who had achieved good academic results in January but had to leave the country some

time later, STRT would have assigned one false negative, and ITNA three. Deducing any sort of trend from the candidates who failed the language test would be careless. The sample sizes are simply too small. Nevertheless, false negatives should be avoided – if only from a social justice perspective.

It is likely that simply raising the overall level of the tests to a C1 level could also raise the level of false negatives. If only candidates with a C1 level on STRT's linguistic criteria had been allowed to register for university STRT would have assigned seven false negatives. This hypothesis is based on scores for written and spoken production. It remains to be seen what the effect would be of using more authentic listening and reading prompts. Nevertheless, given the available information, there is no immediate reason to assume that the cut off score of STRT or ITNA would be inappropriate or substantially misguided. Further research could aim to flesh out the appropriateness of the pass/fail boundary.

EMPIRICALLY ESTABLISH TO WHAT EXTENT STRT AND ITNA SCORES CAN BE CONSIDERED EQUIVALENT

Summary of the findings

Relying on the scores of 118 participants who took STRT and ITNA within the same week, the study discussed in Chapter 3 showed that the correlation between the STRT and ITNA overall and writing scores was moderately high (overall r = .767**; writing r = .694**). The agreement between the scores on the oral tests was much lower however (τ = .387**). Additional analyses revealed further discrepancies. *T*-tests confirmed that the differences between STRT and ITNA mean scores were significant (p < 0.001), with effect sizes ranging from d = -0.53 (writing components) to d = -1.41 (speaking component). Additionally, the pass probability was found to be significantly (p = .02) larger for STRT (.50) than for ITNA (.35). Linear regression and Multi-Faceted Rasch analyses indicated that ITNA's reliance on comparatively difficult *language-in-use* (i.e., grammar and vocabulary) tasks, combined with the inclusion of two relatively easy argumentative tasks in STRT may explain the discrepancy in the written modality. Additionally, Rasch analysis showed that ITNA's spoken component is more difficult than STRT's – again because the relative weight of comparatively difficult linguistic criteria (grammar and vocabulary) in the former, and the relative importance of comparatively easy criteria (e.g., content criteria).

The following chapter considered the scores on the oral components of STRT and ITNA. These components are very similar in terms of task type and rating criteria. Both tests include five criteria that are based on the same CEFR descriptors. Linear and multiple regression and Multi-Faceted Rasch showed that for every CEFR-based criterion ITNA and STRT interpreted the B2 level in a

different way. Weighted kappa coefficients were low for every corresponding criterion ($k_w \leq .216$), and corresponding criteria were never included within the same difficulty band of the multi-faceted Rasch analysis.

Discussion

This study did not find evidence to support that STRT and ITNA operationalize corresponding CEFR-based criteria in comparable tasks in a comparable way. The diverging outcomes can likely be explained, at least in part, by considering the tests' constructs, as well as their differing interpretations of CEFR criteria. ITNA prioritizes lexis and grammar. STRT assigns substantial importance to non-linguistic content criteria, which considerably impacts STRT scores. The fact that there is a significant difference between STRT and ITNA in pass probability, and that one in four candidates receive a different pass/fail judgment should not necessarily be seen as a problem. As long as both tests have been linked to the CEFR in an appropriate way (which appears to be the case), and as long as the validity argument for both tests is convincing for the target context (which does not fully appear to be the case), tests do not need to have the same cut off level. As long as both tests reliably measure at some point within the B2 range, there is no cause for concern from the perspective of the test developers.

From the perspective of the university accepting both tests in a legally equivalent way, a substantial difference in pass/fail judgments, combined with evidence of different interpretations of homonymous constructs and criteria may be cause for concern. It seems opportune for university admission officers to be aware of the differences between STRT and ITNA. Possibly, the university could consider informing prospective students of these differences too. As there is no reason to assume that any test is a better predictor of academic success, students' access to university should not hinge upon the test they chose.

DETERMINE WHETHER ALL STUDENTS WHO ENTER UNIVERSITY WITH A FLEMISH HIGH SCHOOL DEGREE PASS THE B2 THRESHOLD

Summary of the findings

To examine whether all students with a Flemish high school degree have attained the B2 level in Dutch, 159 first-year Flemish L1 students sat two written STRT tasks during their first month of university education. Using non-parametric statistics and Multi-Faceted Rasch analysis, the L1 scores were compared against the performance of two groups of L2 candidates: L2 students who had studied

Dutch at their home institution (N = 629), and L2 students who had done so in Flanders (N = 116). The results showed that L1 students significantly (p < .000) outperformed both groups of L2 students – both overall and on the linguistic criteria of both tasks (when using a conglomerate score for all linguistic criteria used in one task), with medium effect sizes. L2 students who had studied Dutch abroad achieved significantly (p < .000) higher scores on content criteria. Importantly, however, eleven percent of the L1 students did not attain the B2 level as measured by the STRT writing tasks. Logistic regression showed that out of all linguistic criteria, scores on *Grammar* and *Vocabulary* were the best predictors of membership to the group of Flemish students.

Discussion

The fourth chapter did not offer empirical support to the assumption that all Flemish high school graduates possess the B2 level. Logically, it also undermines the assumption that people who have spent one year at a Flemish high school – without necessarily graduating from one – meet the B2 requirement. From a research perspective, this is not entirely surprising. It helps confirm Hulstijn's hypothesis that L2 learners may outperform L1 users on cognitively demanding tasks (Hulstijn, 2015; but see also Stricker, 2004), and corroborates previous findings concerning the importance of vocabulary (Weigle, 2002; Wolfe, Song, & Jiao, 2016) and grammar (di Gennaro, 2016) in distinguishing L1 and L2 performance.

This study – perhaps for the first time – compared L1 users' performance on a centralized L2 test with the performance of L2 learners. The results show quite clearly that graduating from secondary school does not automatically guarantee B2 ability in the language of instruction. Additionally, the results of this study indicate that studying a language in the target context does not automatically lead to higher test scores compared to studying a language in a home context. The implications of both findings will be discussed in Chapter 9.

TRACK AND EXPLAIN LANGUAGE GAINS MADE BY INTERNATIONAL L2 STUDENTS DURING THEIR FIRST YEAR

Summary of the findings

In order to measure the language gains and document the experiences of international L2 students at Flemish universities, twenty respondents were regularly interviewed during their first academic year at a Flemish university. After eight months, the respondents who had not left university (n = 13) took STRT writing and speaking tasks again (the combination of these tasks was predictive for the overall score at R^2_{adj} = .908, p < .000).

The results showed that the respondents had made no significant gains in terms of test score, or in terms of measures of complexity, accuracy, or fluency. The only significant (p = .03) difference was a decreased amount of words used in the oral presentation task. The interview data showed that nearly all respondents had experienced some degree of social and academic isolation, and reported a perceived lack of institutional support. Likely, an important reason why the respondents had made limited or zero gains, was limited opportunities to engage in meaningful interaction with L1 speakers.

Discussion

By operationalizing the Douglas Fir framework, this study has found that identifying the dynamic interactions between institutional mechanisms and interpersonal relationships can help to pinpoint patterns that affect language learning. While it is important to recognize that these findings stem from observations at Flemish universities, it is likely that similar dynamics occur in other contexts as well, in view of the fact that some of the results of this study have been confirmed by previous research. Limited language gains by international students have been reported in Australia (Knoch, Rouhshad, Oon, & Storch, 2015; Knoch, Rouhshad, & Storch, 2014; Storch, 2009); international students' frustrations with didactic traditions were observed in the US, in Canada and in the UK (Amuzie & Winke, 2009; Morita, 2004; Gu, 2005; Gu & Maley, 2008); feelings of alienation, distance, and limited interaction with the L1 community have been reported in the US, Russia, France, Canada, and the UK (Amuzie & Winke, 2009; Gu & Maley, 2008; Kinginger, 2004; Kinginger, 2008; Morita, 2004; Pellegrino Aveni, 2005; Ranta & Meckelborg, 2013). This study thus confirmed previous work, yet also adds to the literature by connecting the dots all the way from language gains to ideology. What the results of this study imply in terms of policy will be discussed below.

CHAPTER 9
LIMITATIONS, IMPLICATIONS & RECOMMENDATIONS

This last section is concerned with the shortcomings and strengths of the research, and with its real-world implications. It brings together the empirical findings to reassess the original assumptions that drove the research questions, and considers how the findings could impact the university admission policy, given the pragmatic reality of policy making.

A FEW WORDS ON STRENGTHS AND LIMITATIONS

Context

This study was conducted in Flanders – an atypical setting in the Anglo-American dominated research domain of university entrance language testing. As is the case with much contextualized research, generalizing the results beyond the original context should be done with utmost care. Even within Flanders, it is quite likely that the results would have been somewhat different had the data been collected at university colleges rather than at universities.

On the other hand, conducting research in the comparatively small and sparsely researched context of Flemish universities has a number of benefits. First, given the educational landscape in Flanders, it was possible to talk with the right policy makers at every university and within the Flemish Department of Education. Additionally, as was discussed in Chapter 6, the modest status of Dutch as an international language ensured that most L2 respondents had not been exposed to it before they had started studying it. In relation to this point, the L2 respondents in Flanders were less likely to use Dutch as a lingua franca when communicating with speakers of a different L1 than they would with a truly international language such as English.

Real-world setting

The rather practical nature of the research goals meant that research data had to be collected in a real-world context (Horii, 2015). This is both a strength and a weakness of the current research. On the one hand, collecting data in the target context eliminates the need to make the kinds of inferences that laboratory-based research would need to make. L2 respondents who took STRT and ITNA bore the real-world consequences of their performance, and tracing those was one of the primary research goals. On the other hand, in the real world,

unforeseen or uncontrollable situations occur, which impacts the data collection in various ways. Some of these factors contributed to the richness of the data (e.g., respondents dropping out offered an insight into their reasons for doing so), but others not so much (e.g., being tied to the May - October period for collecting STRT and ITNA test data impacted the sample size). It is also worth noting that, in contrast to the L2 participants, the test scores did not have real-world consequences for the Flemish students who took two STRT tasks. This may have affected the way both groups approached the tests, and may have impacted their performance.

Lastly, the policy makers, language testers and academic staff were interviewed about a state of affairs that was true at the time of data collection. The results presented here were accurate at the time of data collection but policies may change from year to year. Therefore, the final interviews with policy makers were organized at the end of the writing process. As such we can assume that their testimonies remain factually accurate for at least the remainder of the 2016-2017 academic year.

Population and sample size

The sample sizes used in this research fluctuated quite a bit. The language gains were measured among a relatively small population of thirteen respondents, while the $L2_I$ population consisted of 526 candidates. Although care was taken to interpret the results in the light of sample sizes, and results were backed by referencing other studies, it is important to keep in mind sample sizes when considering score gains, group differences, and effect sizes.

In some cases, population sizes were dictated by practical constraints – a consequence of the real-life setting of this research. The recruiting among the $L2_F$ population could not extend beyond the predetermined natural stopping criterion; the start of the academic year. In other cases, larger group sizes were impossible given the design of the study. In Chapter 2 and Chapter 6, quantitative results were used to add an interpretative layer to qualitative data. The number of participants involved in these studies (N = 55 in Chapter 2, N = 20 in Chapter 6) was rather small for quantitative analyses, but rather substantial given the nature of the qualitative data gathered in the longitudinal design.

Even though the demographics of the sample populations are representative for the actual test population, the sampling methodology used in this study could qualify as convenience sampling. This links in with the real-world nature of the data collected, and may impact the generalizability of the results in Chapters 3 and 4. Because ITNA regulations only allow candidates who passed the written component to take the oral exam, the number of respondents that could meaningfully be compared for the oral component was reduced. Although all statistical assumptions were checked prior to the analyses (Purpura,

Brown, & Schoonen, 2015), range restriction may have had an effect on the data, weakening the correlations.

The act of conducting research

The very act of conducting research influences reality. Undoubtedly, frequently interviewing the $L2_F$ participants during their first academic year at a Flemish university changed the way they experienced that year. During the retrospective interview, various participants explained why being part of this study was a strengthening experience. Even though the researcher never arranged for the participants to meet, some acknowledged that being in the study and knowing that there were other people in the same situation offered them some comfort. Other students felt supported by being able to share their story.

Every $L2_F$ participant in the longitudinal study had a very different story to share, and this resulted in a large dataset. In this dissertation I have looked for patterns and correspondences to link their stories (Chapters 2 and 6). While I have taken care to include important nuances and highlight salient deviations from otherwise consistent observations, no research paper can do justice to each individual story. In order to make sure that no participant felt misrepresented, they were all given insight into the research results based on the interviews, with the explicit invitation to contact me if they wanted a section to be revised. All participants but one were happy with the rendition of the interviews. The one participant who asked for an edit requested that some quotes were omitted, since they were too personal. Needless to say, these quotes were removed from the text.

IMPLICATIONS

The discussion of the policy implications will rely on Fischer's approach to policy evaluation (Fischer, 2003, 2007). Drawing on Toulmin's argumentative structure, Fischer assigns substantial importance to data from mixed methods research, and to a dialogue with policy makers. The first two policy evaluation levels are concerned with policy itself. The two remaining levels focus on the impact of a policy on a wider society. Below, the levels are discussed in pairs.

Policy implications

Program verification involves determining to what extent the policy measures are effective in achieving the policy goal. *Situational validation* is concerned with the assumptions behind the policy, and asks whether every policy measure is equally relevant to solving the perceived problem.

The policy goal, as voiced by the policy makers is to select students who have enough language proficiency to be able to attend a Dutch-medium university program. The perceived problem is students attending class without sufficient language proficiency for doing so successfully. By the policy makers' own assessment, not all policy measures are equally effective in light of the policy goal, but they do represent a workable compromise. Based on the research findings discussed above, Toulmin's argumentative structure can be used to revise the original assumptions.

Assumption 1

The first assumption guiding the research was "B2 is an adequate threshold level to determine international L2 students' access to a Dutch-medium university in Flanders".
Based on (data):

- The interviews with European test developers (Chapter 1), which stress the generally weak empirical foundation for using the B2 level as a university entrance level;
- The academic respondents (Chapter 2), who doubted the adequacy of the B2 level for receptive skills;
- The L2F respondents (Chapter 2), who passed STRT and/or STRT at the B2 level, yet struggled with the linguistic demands of university;
- The very weak relationship between passing a B2 test and achieving academic success (Chapter 2);
- The interviews with policy makers (Chapter 7);

Assumption 1 can be revised into:
B2 *may* (qualifier) function as a minimum threshold for university admission
if (rebuttal 1) the admission officer is aware of the wide performance range inherent in the B2 level,
even though (rebuttal 2) differentiated language level requirements correspond more closely with reality and research consensus.

The warrant connecting the data to the assumption relies on the interpretation and analysis of qualitative data, on the interpretation of non-parametric statistics and on the principles of mixed methods data triangulation. Backing for this warrant can be found in the needs analysis (e.g., Gilabert, 2005; Long, 2005) and predictive validity (e.g., Cho & Bridgeman, 2012; Lee & Greene, 2007) literature.

Assumption 2

The second assumption ("STRT and ITNA are representative of the academic language requirements at Flemish universities") pertains to test development rather than to policy. Even though this is not an assumption for policy makers to justify, it is one they should be informed about.

Based on (data):

- The interviews with the academic respondents and the international L2 students (Chapter 2);
- The field notes, including the class transcriptions (Chapter 2);
- The comparison between essential skills and test task requirements (Conclusion, p. 195);

Assumption 2 can be revised into:
To an extent (qualifier), STRT and ITNA include essential language skills for the target context

but (rebuttal 1)	they sometimes appear to align more with the target *level* than with the target *context*,
and (rebuttal 2)	construct over/underrepresentation does occur.

The warrant combines principles of mixed methods data triangulation and principles of TLU analysis, which have been used in the LSP/LAP literature (Douglas, 2000; Hyland, 2001 Weigle & Malone, 2016), and in the validation literature (Kane, 2017).

Assumption 3

The original formulation of the third assumption was: "STRT and ITNA can be considered equivalent measures of Dutch language proficiency at the B2 level".

Based on (data):

- The level equivalence data (i.e., *t* test, Mc Nemar's test, MFRA, Chapter 3);
- The construct equivalence data (i.e., linear regression, MFRA, Chapter 3);
- The criterion equivalence data (i.e., linear regression, MFRA, Chapter 4);
- The interviews with policy makers (Chapter 6);

Assumption 3 can be revised into:
STRT and ITNA will *likely* (qualifier) assign a corresponding pass/fail judgment to most candidates

but (rebuttal 1)	they operationalize homonymous constructs and criteria differently,
and (rebuttal 2)	may yield a different pass/fail judgment for a substantial proportion of the population around the cut-off score.

The warrant relies on the interpretation of inferential statistics and Multi-Faceted Rasch analysis, which have been used in equivalence research (e.g., ETS, 2010; Riazi, 2013…), concurrent validation (e.g., Lissitz & Samuelsen, 2007…), CEFR-based rating scale design, (e.g., Galaczi et al., 2011; Harsch & Martin 2011…).

Assumption 4

The assumption that "students with a Flemish high school degree have obtained Dutch language proficiency at B2 level" can be rephrased based on the STRT writing test data (Chapter 5):

Most Dutch-medium high school graduates (typically from the academically-oriented strand) will *likely* (qualifier) meet the B2 threshold for writing

but (rebuttal 1)	most probably this claim cannot be maintained for students with a different home language or an atypical SE program.

The warrant relies on the interpretation of non-parametric statistics and Multi-Faceted Rasch analysis, which have been used in – among others – L1/L2 writing research (e.g., Leki et al., 2008; Polio, 2013; Weigle & Friginal, 2015).

Assumption 5

Based on the research data, the original formulation of Assumption 5 ("International L2 students will make language gains by virtue of studying at a Flemish university") cannot be maintained.

Based on the analysis of the STRT test/retest data and on the longitudinal interviews, it is clear that: "International L2 students in Flanders who attend large-scale study programs must not be assumed to make productive language gains if there is no appropriate post-admittance policy in place".

Since this is not an assumption but a statement, there are no rebuttals. The warrant linking the data to the statement relies on non-parametric statistics (used to measure differences in terms of score, linguistic complexity, accuracy, or

fluency) and on principles of mixed-method data triangulation, previously employed in the L2 language gains, and the study abroad literature (e.g., Storch, 2009; Knoch, 2015; Serrano et al., 2012). Assumption 5 does not directly impact the admission requirements, but does have implications for the post-admission policy, which will be discussed below.

In sum, the Flemish university admission policy does not appear to be fully effective to meet the policy goal: Many students who enter university will still experience linguistic problems. Additionally, it has unwanted side-effects (such as false negatives), it includes empirically unjustified exemptions, and it relies on partially unfounded assumptions. At the same time these policy characteristics do not appear to deviate substantially from the average European university admission regulations. Moreover, the limited effectiveness of the university admission policy in Flanders does not contrast with the policy makers' assessment, who are fully aware that some admission requirements are concessions to or compromises with important stakeholders.

The revised assumptions are the result of academic rather than pragmatic reasoning, and as such they may be too nuanced to be useable for policy makers. They have, however been used as the foundation for concrete policy recommendations, at the end of this chapter.

Societal implications

The societal layer of Fischer's policy evaluation model incorporates the level of *societal vindication*, which examines whether a policy contributes positively to the wider context in which it operates, and the level of *social choice*, which is concerned with the ideology on which a policy is based, and questions whether that ideology would be conducive to building a society that embraces equality and freedom. The basic concerns of *societal vindication* and *social choice* align well with the definition of justice operationalized in Chapter 2.

Ideologically, the university language regulations show a tendency for linguistic protectionism (see Introduction and Chapter 5): There are strict language quota, and legally binding language requirements for international professors. Likewise, every university has chosen to implement language requirements for international students. As was argued in the introduction, the language ideology in Flanders could be described as territorial monolingualism. Central to this idea is the conception that Flemish L1 users of Dutch are the norm, and international students or migrants are expected to adjust to that norm, down to the level of pronunciation (Blommaert, 2011; Van Splunder, 2016). In such a context, introducing language tests for university admission is not surprising. It is far from unique however: University entrance is one of the primary fields of high-stakes language testing worldwide. At the same time, the

justice of having tests determine access to education is rarely questioned (McNamara & Ryan, 2011).

Justice authors have argued that the introduction of requirements and regulations that limit the access of one specific subgroup from a larger population may cause inequities (Kunnan, 2000, 2004). In Chapter 2 it was argued, relying on Sen (2010), that a policy is likely to be unjust if it restricts test takers' freedom of access without empirical or reasonable motivation. A policy may thus limit the access of certain people to certain goods, services or capabilities, only if it relies on careful examination of solid evidence. Absence of evidence, or the presence of negative evidence implies that a policy is unjust (Dworkin, 2013). There is of course always a chance that tests or policies discriminate adequately by accident, but as Sen argues, this is insufficient. He uses the analogy of a clock to demonstrate his point (Sen, 2010: 40): A broken clock is exactly right twice a day, but that does not make it more reliable than a watch which runs a little behind. Sen thus prefers a policy that relies on reasoned scrutiny, not because this yields a perfectly just system, but because it implies a concern for an evidence-based policy.

If we apply this logic to the current research, it is clear that the purpose of the Flemish admission policy (i.e., ensuring that incoming international L2 students possess the language proficiency required to attend university) is ethically defensible. Given the number of false positives and false negatives, however, the system in place appears rather ineffective. Additionally, it discriminates among international students (but not *all* international students, since the entrance requirements do not apply to certain programs) based on a criterion that not all Flemish students meet. All in all, the empirical foundation supporting the admission system appears rather thin, and most of the policy makers' assumptions are unsupported by empirical data. Even within the paradigm of territorial monolingualism, this dissertation strongly suggests that the Flemish university admission policy could be revised to meet higher standards in terms of justice and empirical foundation. Evidence from other countries indicated that Flanders is not exceptional in this regard (e.g., Carlsen, 2017) – it is simply the first context in which the primary assumptions have been systematically examined from multiple perspectives in one study.

RECOMMENDATIONS

Recommendations for practice

There are a number of possible ways in which this dissertation could have real-world impact. One possible, albeit undesirable, scenario is zero impact: nothing happens. By including policy makers in this study, and talking to them about this

study, this outcome has hopefully been prevented, but Chapter 7 did show quite clearly that empirical data do not always lead to policy changes. However, if there is willingness to improve current practice – and this seems to be the case – relatively straightforward adjustments by different stakeholders could yield important improvements in the field.

Language test developers could consider aligning the test constructs more with the real-world requirements. Listening tasks could feature more authentic prompts, which may fuel positive washback. The weighting of oral tasks could be reduced in order to align better with the actual importance of speaking at university. These recommendations should not be rushed, of course, but they seem worthy of investigation. In line with this, the admission requirements of universities could be re-examined. Conducting needs analyses in order to draw up language requirements that meet the actual language needs of prospective students would be a major step forward, but given the reality of stakeholders influencing admission requirements for the purpose of controlling student numbers, this will likely not happen soon.

Perhaps the most important recommendation concerns not the entrance requirements, but the post-admittance policy. After international L2 students are registered for Dutch-medium programs, they become part of the general population. No specific accommodations are in place for this group. To recognize the presence of this group (e.g., by welcoming them specifically), to create the circumstances that would help them build a network, and to address their language-related needs (e.g., by providing class recordings), would be a big step forward.

Finally, as Byrnes et al. (2010) argued, the responsibility of a university does not stop at admission – that is when it begins. Consequently, universities could consider setting attainment targets in addition to entrance requirements. Universities could help international L2 students make language gains by providing curricular language classes that offer language support throughout their academic trajectory.

Implications for research

Each chapter lists the research gaps addressed in this dissertation. Briefly summarized, this dissertation has provided data regarding the use of the CEFR in university admission testing, has examined context representativeness from the perspective of justice, has scrutinized test equivalence by considering item-level scores and rating scale descriptors, has compared L1 and L2 test performance, has offered an explanation for limited spoken and written language gains by considering the social context, and has traced the origins of the admission policy by consulting policy makers. This is, to the best of my knowledge, the first study

to systematically examine the main assumptions behind an admission policy from a variety of perspectives.

Methodologically, this dissertation has proposed a number of innovations, such as bypassing truncated sample issues, operationalizing level and construct equivalence, estimating rating scale descriptor similarity by using the Jaccard Index, and utilizing the Douglas Fir framework as a method of qualitative analysis. Undoubtedly, there are limitations to this research (see above) and its operationalization, but future research could use these approaches as a starting point for further examination.

The following three implications could serve as an inspiration for additional research. First, this dissertation demonstrates the need to keep examining and validating important test-related assumptions – no matter how widely-held they are (Connolly, Arkes, & Hammond, 1999; McNamara & Ryan, 2011; Phillips, 2007). Second, this research echoes that test validation can and should go beyond the test itself, and consider the impact on society (Kane, 2013; Messick, 1989), not only from a rational, but also from a reasonable perspective (Toulmin, 2001). Thirdly, related to the idea of reasonableness, it is important for language testers to engage with real-world practice. If policy makers were more actively involved in research, or if they were more actively informed about research, maybe our research would have more of an impact, and maybe our recommendations would have more resonance.

"THE NEED FOR VALIDATION, BABY GONE COMPLETELY BERSERK"

Quite likely, Nick Cave, in the motto of this dissertation, referred to a different type of validation than Messick or Kane. But was he right, anyway? I would argue that the *need* for validation is perhaps more pressing now than it has ever been.

In an age of unprecedented migration flows, more people – refugees, students, academics – are affected by the use of language requirements as gatekeepers to valued statuses, goods, or services. As this dissertation has shown, the motivation for these requirements may reside in empirically unsound assumptions or unverified claims. Logically, as the number of people impacted rises, so does the extent of the impact of misguided assumptions and populist claims.

This evolution emphasizes the social responsibility of language testers. If it is the purpose of science to find truth, and the purpose of policy to advance justice, then certainly there is a role for research to hold policy claims to the light, and distinguish between truth and untruth.

REFERENCES

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

ACTFL. (2016). *Assigning CEFR Ratings to ACTFL Assessments. American Council On The Teaching Of Foreign Languages*. Retrieved October 20, 2015, from www.actfl.org.

Agirdag, O. (2010). Exploring bilingualism in a monolingual school system: insights from Turkish and native students from Belgian schools. *British Journal of Sociology of Education*, 31(3), 307–321.

Alderson, C. (1991). Bands and scores. In C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). London: Macmillan.

Alderson, C. (2007). The CEFR and the Need for More Research. The Modern Language Journal, 91(4), 659–663.

Alderson, C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project. *Language Assessment Quarterly*: An International Journal, 3(1), 3-30.

ALTE. (2001). *Principles of good practice for ALTE examinations*. Retrieved February 13, 2014, from www.alte.org.

Amkreutz, R. (2013, September 24). *Amper 40 procent van de eerstejaarsstudenten slaagt*. Retrieved May 25, 2016, from www.demorgen.be.

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

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing*, 18(3), 191–208.

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

Bachman, L. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, 19(3), 535–556.

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing*, 19(4), 453–476.

Bachman, L., Davidson, F., & Ryan, K. (1995). *An Investigation Into the Comparability of Two Tests of English as a Foreign Language*. Cambridge: Cambridge University Press.

Ball, S. (2015). What is policy? 21 years later: reflections on the possibilities of policy research. *Discourse: Studies in the Cultural Politics of Education*, 36(3), 306–313.

Ball, S., Maguire, M., & Braun, A. (2012). *How Schools Do Policy: Policy Enactments in Secondary Schools*. New York: Routledge.

Bärenfänger, O., & Tschirner, E. (2012). *Assessing Evidence of Validity of Assigning CEFR Ratings to the ACTFL Oral Proficiency Interview (OPI) and the Oral Proficiency Interview by computer (OPIc)*. Leipzig: Institute for Test Research and Test Development.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, 18(3), 279–93.

Barkaoui, K. (2014). Multifaceted Rasch Analysis for Test Evaluation. In A. Kunnan (Ed.) *The Companion to Language Assessment* (pp. 1301–1322.). New Jersey: John Wiley & Sons, Inc.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), 497–529.

Baynham, M. (2006). Performing self, family and community in Moroccan narratives of migration and settlement. In A. Da Fina, D. Schiffrin, & M. Bamberg (Eds.), *Discourse and Identity* (pp. 376–397). Cambridge: Cambridge University Press.

Baztán, A. (2008). *La evaluación oral: una equivalencia entre las guidelines de ACTFL y algunas escalas del MCER*. Granada: Universidad de Granada.

Beleidscel Diversiteit en Gender. (2016). Diversiteit aan de UGent: de instroom van kansengroepen in cijfers. Ghent University, unpublished policy document.

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 Publishing.

Block, D. (2007). The Rise of Identity in SLA Research, Post Firth and Wagner (1997). *The Modern Language Journal*, 91, 863–876.

Blommaert, J. (2011). The long language-ideological debate in Belgium. *Journal of Multicultural Discourses*, 6(3), 241–256.

Blommaert, J., & Van Avermaet, P. (2008). *Taal, onderwijs en de samenleving. De kloof tussen beleid en realiteit*. Berchem: Epo.

Bond, T., & Fox, C. (2007). Applying the Rasch Model: *Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Lawrence Erlbaum.

Borg, S. (2006). *Teacher Cognition and Language Education: Research and Practice*. London: Continuum.

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

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.

Bouma, K. (2016, August 26). Meer dan de helft van de studies volledig in het Engels. *De Volkskrant*. Retrieved on November 14, 2016, from www.volkskrant.nl.

Bourdieu, P. (1991). *Language and Symbolic Power*. Cambridge, Mass: Harvard University Press.

Bovens, M., 't Hart, P., & Kuipers, S. (2006). The politics of policy evaluation. In M. Moran, M. Rein, & R. E. Goodin (Eds.), *The Oxford handbook of public policy* (pp. 319–336). Oxford: Oxford University Press.

Braine, G. (2002). Academic literacy and the nonnative speaker graduate student. *Journal of English for Academic Purposes*, 1(1), 59–68.

Buckley, K., Winkel, R., & Leary, M. (2004). Reactions to acceptance and rejection: Effects of level and sequence of relational evaluation. *Journal of Experimental Social Psychology*, 40(1), 14–28.

Byrnes, H. (2007). Perspectives. *The Modern Language Journal*. 91: 641–5.

Byrnes, H., Maxim, H., & Norris, J. (2010). Realizing Advanced Foreign Language Writing. *The Modern Language Journal*, 94, 1–202.

Cai, H. (2013). Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing*, 30(2), 177–199.

Carlsen, C.H. (2017, in press). The adequacy of the B2-level as university entrance requirement. *Language Assessment Quarterly*.

Carlsen, C.H. (2014). How valid is the CEFR as a construct for language tests? Presented at the *ALTE 5th International Conference*, Paris, France, 10-11 April 2014.

Cave, N. (2004). Abattoir Blues. *Abattoir Blues / The Lyre of Orpheus*. London: Mute Records.

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, 26, 20–37.

Chirkov, V., Vansteenkiste, M., Tao, R., & Lynch, M. (2007). The role of self-determined motivation and goals for study abroad in the adaptation of international students. *International Journal of Intercultural Relations*, 31, 199–222.

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.

CNaVT. (2013). *Jaarverslag 01/09/2011 - 31/08/2012*. CNaVT/Nederlandse Taalunie, unpublished policy document.

CNaVT. (2014). *STRT Validity Argument*. CNaVT/Nederlandse Taalunie, unpublished policy document.

CNaVT. (2016a). *Educatief Startbekwaam*. Retrieved October 25, 2016, from www.cnavt.org

CNaVT. (2016b). *Jaarverslag 2015*. CNaVT/Nederlandse Taalunie, unpublished policy document.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Earlbaum Associates.

Collentine, J., & Freed, B. (2004). Learning context and its effects on second language acquisition. Studies in Second Language Acquisition, 26, 153–171.

Connolly, T., Arkes, H., & Hammond, K. (1999). *Judgment and Decision Making: An Interdisciplinary Reader*. Cambridge: Cambridge University Press.

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

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

Cumming, A. (2013). Assessing Integrated Writing Tasks for Academic Purposes: Promises and Perils. *Language Assessment Quarterly*, 10(1), 1–8.

Curtis, A. (2015). Discussant at the ILTA-AAAL joint event: *Revisiting the Interfaces between SLA and Language Assessment Research*. Toronto, Canada, 21-24 March 2015.

Davies, A. (1984). Validating three tests of English language proficiency. *Language Testing*, 1(1), 50–69.

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

De Beer, J., Smith, U., & Jansen, C. (2009). "Situated" in a separated campus – Students' sense of belonging and academic performance: A case study of the experiences of students during a higher education merger. *Education as Change*, 13(1), 167–194.

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

De Costa, P. I. (2010). Language ideologies and standard English language policy in Singapore: responses of a "designer immigrant" student. *Language Policy*, 9(3), 217–239.

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

De Jong, J. (2013). How to save the Common European Framework. Presented at *Language Testing in Europe: Time for a new Framework*? Antwerp.

De Jong, J. (2014). Standards & Scaling. Presented at *LTRC*, Amsterdam.

De Jong, J., Becker, K., Bolt, D, & Goodman, J. (2014). *Aligning PTE Academic Test Scores to the Common European Framework of Reference for Languages*. Retrieved on July 8, 2015, from http://pearsonpte.com.

de Larios, J. R., Marín, J., & Murphy, L. (2001). A Temporal Analysis of Formulation Processes in L1 and L2 Writing. *Language Learning*, 51(3), 497–538.

De Standaard. (2013). Selecteer studenten na eerste semester. Retrieved on July 3, 2014, from www.standaard.be.

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

De Wachter, L., Heeren, J., Marx, S., & Huyghe, S. (2013). Taal: noodzakelijke, maar niet enige voorwaarde tot studiesucces. Correlatie tussen resultaten van een taalvaardigheidstoets en slaagcijfers bij eerstejaarsstudenten aan de KU Leuven. *Levende Talen Tijdschrift*, 14(4), 28–36.

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

Dehandschutter, W. (2016). Een app om anoniem de professor wat meer uitleg te vragen. De Standaard. Retrieved on March 6, 2016, from www.standaard.be.

Departement Onderwijs en Vorming. (2015). *Taalverslag academiejaar 2013-2014. Departement Onderwijs en Vorming. Afdeling Hoger Onderwijs en Volwassenenonderwijs*. Retrieved from on January 4, 2016, from www.vlaanderen.be.

Departement Onderwijs en Vorming. (2016). *Screening niveau onderwijstaal, taaltraject en taalbad in het gewoon lager onderwijs.* BaO/2014/01. Retrieved on May 24th, 2016, from http://data-onderwijs.vlaanderen.be.

Dewey, D. P., Bown, J., Baker, W., Martinsen, R. A., 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.

Deygers, B., & Luyten, L. (2012). *Verslag ITNA/PTHO---onderzoeksoverleg 2011-2012.* Unpublished policy document.

Deygers, B., De Wachter, L., Van Gorp, K., & Joos, S. (2013). Examining the concurrent validity of a task-based and an indirect academic language test. Presented at TBLT, Banff, Canada.

Deygers, B., Van Gorp, K., Luyten, L., & Joos, S. (2013). Rating scale design: a comparative study of two analytic rating scales in a task-based test. In E. Galaczi & C. Weir (Eds.), Exploring Language Frameworks. *Proceedings from the ALTE Kraków Conference*. (Vol. 36, pp. 273–289). Cambridge: Cambridge University Press.

di Gennaro, K. (2009). Investigating differences in the writing performance of international and Generation 1.5 students. *Language Testing*, 26(4), 533–559.

di Gennaro, K. (2013). How different are they? A comparison of Generation 1.5 and international L2 learners' writing ability. *Assessing Writing*, 18(2), 154–172.

di Gennaro, K. (2016). Searching for differences and discovering similarities: Why international and resident second-language learners' grammatical errors cannot serve as a proxy for placement into writing courses. *Assessing Writing*, 29, 1–14.

Díaz-campos, M. (2004). Context of Learning in the Acquisition of Spanish Second Language Phonology. *Studies in Second Language Acquisition*, 26(2), 249–273.

Duff, P. A. (2002). The Discursive Co-construction of Knowledge, Identity, and Difference: An Ethnography of Communication in the High School Mainstream. *Applied Linguistics*, 23(3), 289–322.

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

Dworkin, R. (2013). *Justice for Hedgehogs* (Reprint edition). Cambridge, Mass: Belknap Press.

EACEA (2010). *Focus on Higher Education in Europe*. Brussels: Education, Audiovisual and Culture Executive Agency.

EALTA. (2000). EALTA *Guidelines for good practice in language testing and assessment*. Retrieved on April 16, 2014, from www.ealta.eu.org.

EEA. (2001). *Reporting on environmental measures: Are we being effective?* Copenhagen: European Environment Agency.

Elder, C., & O'Loughlin, K. (2003). Investigating the Relationship between Intensive English Language Study and Band Score Gain on IELTS. *IELTS Research Reports*, 4, 207–254.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Engle, L., & Engle, J. (2003). Study abroad levels: Toward a classification of program types. *Frontiers: The Interdisciplinary Journal of Study Abroad*, 9, 1-20.

ETS. 2010. Linking *TOEFL iBTTM Scores to IELTS® Scores—A Research Report*. Princeton: Educational Testing Service.

Field, A, Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. London: Sage.

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

Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, 66(4), 477–485.

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*. Stasbourg: Council of Europe.

Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: a manual. *Language Testing*, 22(3), 261–279.

Fischer, F. (2007). Deliberative policy analysis as practical reason: integrating empirical and normative arguments. In F. Fischer & G. J. Miller (Eds.), *Handbook of Public Policy Analysis: Theory, Politics, and Methods* (pp. 223–236). Boca Raton: CRC Press.

Fløttum, K., Gedde-Dahl, T., & Kinn, T. (2006). *Academic Voices: Across Languages and Disciplines*. Amsterdam: John Benjamins Publishing.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics*, 21(3), 354–375.

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

Fox, J. (2005). Rethinking second language admission requirements: problems with language-residency criteria and the need for language assessment and support. *Language Assessment Quarterly*, 2(2), 85–115.

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

Fulcher, G. (1997). An English Language Placement Test: Issues in Reliability and Validity. *Language Testing* 14, 113–39

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

Fulcher, G. (2012a). Assessment Literacy for the Language Classroom. *Language Assessment Quarterly*, 9(2), 113–132.

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

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

Galaczi, E., 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.

Gass, S. (2003). Input and interaction. In C. Doughty & M. Long (Eds.), *Handbook of second language acquisition* (pp. 224–250). Oxford: Blackwell.

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.

Ginther, A., & Stevens, J. (1998). Language background, ethnicity, and the internal construct validity of the Advanced Placement Spanish Language Examination. In A. J. Kunnan (Ed.), *Validation in language assessment*. (pp. 169–194). Mahwah, NJ: Lawrence Erlbaum.

Glaser, B. & Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine de Gruyter.

Glorieux, I., Laurijssen, I., & Sobczyk. (2015). *Studiesucces in het eerste jaar hoger onderwijs in Vlaanderen. Een analyse van de impact van kenmerken van studenten en van opleidingen*. Leuven: Steunpunt SSL.

Gogolin, I. (2002). Linguistic and Cultural Diversity in Europe: A Challenge for Educational Research and Practice. *European Educational Research Journal*, 1(1), 123–138.

Gomez, P., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, 24(3), 417–444.

Goodin, R. E., Rein, M., & Moran, M. (2006). The public and its policies. In M. Moran, M. Rein, & R. E. Goodin (Eds.), *The Oxford handbook of public policy* (pp. 3–39). Oxford: Oxford University Press.

Goovaerts, M. (2012). *Wie overleeft het eerste bachelorjaar niet? Een onderzoek naar drop-out in het hoger onderwijs*. Brussel: Vlaams Verbond Van Katholieke Hogescholen.

Gorin, J. (2007). Reconsidering Issues in Validity Theory. *Educational Researcher*, 36(8), 456–462.

Gorter, D., & Cenoz, J. (2012). Regional minorities, education and language revitalization. In M. Martin-Jones, A. Blackledge, & A. Creese (Eds.), *The Routledge Handbook of Multilingualism* (pp. 184–198). New York: Routledge.

Green, A. (2004). Making the grade: score gains on the IELTS Writing test. *Cambridge ESOL Research Notes*, 16, 9–13.

Green, A. (2017, in press). Linking Tests of English for Academic Purposes to the CEFR: The Score User's Perspective. *Language Assessment Quarterly*.

Grin, F. (2000). *Evaluating policy measures for minority languages in Europe: towards effective, cost-effective and democratic implementation*. Flensburg: European Centre for Minority Issues.

Grin, F. (2003). *Language policy evaluation and the European charter for regional or minority languages*. New York: Palgrave Macmillan.

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing*, 31(1), 111–133.

Gu, Q., & Maley, A. (2008). Changing Places: A Study of Chinese Students in the UK. *Language and Intercultural Communication*, 8(4), 224–245.

Gysen, S., & Avermaet, P. V. (2005). Issues in Functional Language Performance Assessment: The Case of the Certificate Dutch as a Foreign Language. *Language Assessment Quarterly*, 2(1), 51–68.

Hamilton, J., Lopes, M., McNamara, T., & Sheridan, E. (1993). Rating scales and native speaker performance on a communicatively oriented EAP test. *Language Testing*, 10(3), 337–353.

Harklau, L., Losey, K. M., & Siegal, M. (1999). *Generation 1.5 Meets College Composition: Issues in the Teaching of Writing To U.S.-Educated Learners of ESL*. Mahwah, N.J: Routledge.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, 17(4), 228–250.

Harsch, C., & Rupp, A. (2011). Designing and Scaling Level-Specific Writing Tasks in Alignment With the CEFR: A Test-Centered Approach. *Language Assessment Quarterly*, 8(1), 1–33.

Hechanova-Alampay, R., Beehr, T. A., Christiansen, N. D., & Van Horn, R. (2002). Adjustment and Strain among Domestic and International Student Sojourners: A Longitudinal Study. *School Psychology International*, 23(4), 458–74.

Herelixka, C. (2013). *Academisch taalvaardig: van start tot finish De invloed van meertaligheid op de TaalVaST-toets*. Leuven, unpublished master thesis.

Hernández, T. A. (2010). Promoting Speaking Proficiency through Motivation and Interaction: The Study Abroad and Classroom Learning Contexts. *Foreign Language Annals*, 43(4), 650–670.

Hill, K., Storch, N., & Lynch, B. (1999). *A Comparison of IELTS and TOEFL as Predictors of Academic Success (Vol. 2)*. Retrieved on September 4, 2012, from www.ielts.org.

Hirvela, A. (2016). Academic reading into writing. In K. Hyland & P. Shaw (Eds.), *The Routledge Handbook of English for Academic Purposes* (pp. 127–139). London and New York: Routledge.

Holmes, J., Marra, M., & Vine, B. (2011). *Leadership, discourse, and ethnicity*. Oxford: Oxford University Press.

Horii, S. Y. (2015). Second language acquisition and language teacher education. In M. Bigelow & J. Ennser-Kananen (Eds.), *The Routledge handbook of educational linguistics* (pp. 313–324). New York: Routledge.

Housen, A., Schoonjans, E., Janssens, S., Welcomme, A., Schoonheere, E., & Pierrard, M. (2011). Conceptualizing and measuring the impact of contextual factors in instructed SLA - the role of language prominence. *IRAL: International Review of Applied Linguistics in Language Teaching*, 49(2), 83–112.

Howell, D. (1997). *Statistical methods for psychology*. Belmont, CA: Duxbury.

Howlett, M., & Giest, S. (2013). The policy-making process. In E. J. Aral, S. Fritzen, M. Howlett, M. Ramesh, & X. Wu (Eds.), *Routledge handbook of public policy* (pp. 17–28). London & New York: Routledge.

Huie, K., & Yahya, N. (2003). Learning to Write in the Primary Grades: Experiences of English Language Learners and Mainstream Students. *TESOL Journal*, 12(1), 25–31.

Hulstijn, J. (2007). The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency. *The Modern Language Journal*, 91(4), 663–667.

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.

Hulstijn, J. (2015). *Language Proficiency in Native and Non-native Speakers: Theory and research*. Amsterdam ; Philadelphia: John Benjamins Publishing Company.

Hume, D. (1978, 1738), *A Treatise of Human Nature*. Oxford: Clarendon Press.

Hyland, K. (2016). General and specific EAP. In K. Hyland & P. Shaw (Eds.), *The Routledge Handbook of English for Academic Purposes* (pp. 17–30). London and New York: Routledge.

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, 8 March 2014.

ILTA. (2007). Guidelines for Practice. International Language Testing Association. Retrieved on April 28, 2016, from www.iltaonline.com.

Interuniversitair Testing Consortium. (2015). *Interuniversitaire Taaltest Nederlands voor Anderstaligen. Zelfevaluatierapport*. Unpublished policy document.

ITNA. (2016). *Testprincipes*. Retrieved October 25, 2016, from www.itna.be.

Jann, W., & Fischer, F. (2007). Deliberative policy analysis as practical reason: integrating empirical and normative arguments. In F. Fischer & G. J. Miller (Eds.), *Handbook of Public Policy Analysis: Theory, Politics, and Methods* (pp. 223–236). Boca Raton: CRC Press.

Jann, W., & Wegrich, K. (2007). Theories of the Policy Cycle. In F. Fischer & G. J. Miller (Eds.), *Handbook of Public Policy Analysis: Theory, Politics, and Methods* (pp. 43–62). Boca Raton: CRC Press.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2), 112–133.

Jordens, K. (2016). *Turkish is not for learning, Miss. Valorizing linguistic diversity in primary education*. Leuven: KU Leuven.

Juan-Garau, M., Salazar-Noguera, J., & Prieto-Arranz, J. I. (2014). L2 English learners' lexico-grammatical and motivational development at home and abroad. In C. Pérez-Vidal (Ed.), *Language Acquisition in Study Abroad and Formal Instruction Contexts* (pp. 235–259). Amsterdam: John Benjamins.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus & Giroux.
Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31–36.
Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, 38(4), 319–342.
Kane, M. (2010). Validity and fairness. *Language Testing*, 27(2), 177–182.
Kane, M. T. (2012). Articulating a validity argument. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 34–48). London and New York: Routledge.
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). Routledge.
Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing*, 15(4), 261–278.
Keck, C. (2014). Copying, paraphrasing, and academic writing development: A re-examination of L1 and L2 summarization practices. *Journal of Second Language Writing*, 25, 4–22.
Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. *IELTS Research Reports*, 3, 85–108.
Khalifa, H., & ffrench, A. (2009). Aligning Cambridge ESOL examinations to the CEFR: issues and practice. *Cambridge Research Notes* 37, 10-15.
Khan, K., & McNamara, T. (2017, in press). Citizenship, immigration laws, and language. In S. Canagarajah (Ed.), *The Routledge Handbook of Migration and Language* (pp. 451–467). London & New York: Routledge.
Kinginger, C. (2004). Alice doesn't live here anymore: Foreign language learning and identity construction. In A. Pavlenko & A. Blackledge (Eds.), *Negotiation of identities in multilingual contexts* (pp. 219–242). Clevedon: Multilingual Matters.
Kinginger, C. (2008). Language Learning in Study Abroad: Case Studies of Americans in France. *The Modern Language Journal*, 92, 1–124.
Kinginger, C. (2010). American students abroad: Negotiation of difference? *Language Teaching*, 43(2), 216–227.
Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing*, 21, 1–17.
Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing*, 28, 39–52.
Koda, K. (1993). Task-Induced Variability in FL Composition: Language-Specific Perspectives. *Foreign Language Annals*, 26(3), 332–346.

Kormos, J., Csizér, K., & Iwaniec, J. (2014). A mixed-method study of language-learning motivation and intercultural contact of international students. *Journal of Multilingual and Multicultural Development*, 35(2), 151–166.

Krashen, S. (1985). *The Input Hypothesis: Issues and Implications*. London, New York: Longman.

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.

KU Leuven. (2016, April 28). *Onderwijs- en examenreglement 2016-2017*. Retrieved September 30, 2016, 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. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). 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.

Lantolf, J. P., & Genung, P. B. (2003). "I'd rather switch than fight": An activity-theoretic study of power, success, and failure in a foreign language classroom. In C. Kramsch (Ed.), *Language Acquisition and Language Socialization: Ecological Perspectives* (pp. 175–197). London & New York: Bloomsbury Academic.

Lave, J., & Wenger, E. (1992). *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press.

Lee, E. (2008). The "other(ing)" costs of ESL: A Canadian case study. *Journal of Asian Pacific Communication*, 18(1), 91–108.

Lee, M. Y. P. (2003). Discourse structure and rhetoric of English narratives: Differences between native English and Chinese non-native English writers. *Text*, 23(3), 347.

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.

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.

Leki, I., Cumming, A., & Silva, T. (2008). *A Synthesis of Research on Second Language Writing in English*. New York: Routledge.

Leliaert, S. (2011). België, het beloofde studieland? Retrieved June 15, 2016, from http://www.dewereldmorgen.be.

Leung, C., Harris, R., & Rampton, B. (2004). Living with inelegance in qualitative research on task-based learning. In K. Toohey & B. Norton (Eds.), *Critical pedagogies and language learning* (pp. 242–267). Cambridge: Cambridge University Press.

Lievens, S. (2016). Diversiteit aan de UGent: de instroom van kansengroepen in cijfers. Ghent University, unpublished policy document.

Linacre, J. (2012). *A user's guide to FACETS Rasch-model computer programs*. Retrieved, November 20, 2013, from www.winsteps.com.

Linacre, J. (2015). *FACETS* (Version 3.71.4). Beaverton, Oregon: Winsteps.com.

Lindridge, A. (2015). Construct Equivalence. In C. Cooper (Ed.),*Wiley Encyclopedia of Management* 9, 1–3. New Jersey: Wiley.

Linton, A. (2009). Language politics and policy in the United States: implications for the immigration debate. *International Journal of the Sociology of Language*, (199), 9–37.

Lissitz, R., & Samuelsen, K. (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher*, 36(8), 437–448.

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.

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.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, 19(3), 246–276.

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

MacDonald, G., & Leary, M. R. (2005). Why does social exclusion hurt? The relationship between social and physical pain. *Psychological Bulletin*, 131(2), 202–223.

Maes, C. (2016). Het CNaVT vernieuwt. Retrieved on 2 November, 2016, from www.tijdschriftles.nl.

Maestas, R., Vaquera, G. S., & Zehr, L. M. (2007). Factors Impacting Sense of Belonging at a Hispanic-Serving Institution. *Journal of Hispanic Higher Education*, 6(3), 237–256.

Magnan, S., & Back, M. (2007). Social Interaction and Linguistic Gain During Study Abroad. *Foreign Language Annals*, 40(1), 43–61.

Malone, M. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing*, 30(3), 329–344.

Manderson, L., Bennett, E., & Andajani-Sutjahjo, S. (2006). The Social Dynamics of the Interview: Age, Class, and Gender. *Qualitative Health Research*, 16(10), 1317–1334.

Manning, C., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Martyniuk, W. (2010). *Studies in Language Testing 33: Aligning Tests with the CEFR. Reflections on using the Council of Europe's draft Manual*. Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. (2001). Language assessment as social practice: challenges for research. *Language Testing*, 18(4), 333–349.

McNamara, T. (2007). The CEFR in Europe and Beyond. Paper presented at the 4th Annual EALTA Conference, Sitges, Spain.

McNamara, T. (2012). Language Assessments as Shibboleths: A Poststructuralist Perspective. *Applied Linguistics*, 33(5), 564–581.

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). Washington, DC: American Council on Education/Macmillan.

Miles, M. & Huberman A. (1994). *Qualitative Data Analysis*. Baverly Hills: Sage.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2013). *Qualitative Data Analysis: A Methods Sourcebook*. Thousand Oaks, CA: Sage.

Morita, N. (2004). Negotiating Participation and Identity in Second Language Academic Communities. *TESOL Quarterly*, 38(4), 573–603.

Moss, P. (2007). Reconstructing Validity. *Educational Researcher*, 36(8), 470–476.

Nagel, S. (2002). *Handbook of public policy evaluation*. Thousand Oaks, CA: Sage.

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

Nederlandse Taalunie. (2013). *Aankondiging van een opdracht - Diensten*. Den Haag: Nederlandse Taalunie.

Negishi, M., Takada, T. & Tono, Y. (2013) A progress report on the development of the CEFR-J. In E. Galaczi & C. Weir (Eds.), *Exploring Language Frameworks. Proceedings from the ALTE Kraków Conference*. Cambridge: Cambridge University Press.

Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing*, 19(4), 337–346.

Norris, J.M. (2008). *Validity Evaluation in Language Assessment*. New York: Peter Lang.

Norris, J.M. (2015). Statistical Significance Testing in Second Language Research: Basic Problems and Suggestions for Reform. *Language Learning*, 65, 97–126.

Norris, J.M. (2016). Reframing the SLA-Assessment Interface: 'Constructive' Deliberations at the Nexus of Interpretations, Contexts, and Consequences. Paper presented at the *Language Testing Research Colloquium*, Palermo, Italy.

North, B. (2007). The CEFR Illustrative Descriptor Scales. *The Modern Language Journal*, 91(4), 656 - 659.

North, B. (2014a). *English Profile Studies. The CEFR in Practice*. Cambridge: Cambridge University Press.

North, B. (2014b). Putting the Common European Framework of Reference to good use. *Language Teaching*, 47(2), 228–249.

North, B. (2016). Use and misue of the CEFR in teaching and assessment. Presented at the *ALTE 48th Conference Day*, Stockholm.

Norton, B. (2013). *Identity and Language Learning: Extending the Conversation*. Bristol: Multilingual Matters.

Norton, B., & Toohey, K. (2011). Identity, language learning, and social change. *Language Teaching*, 44(04), 412–446.

Nussbaum, M. (2002). Capabilities and Social Justice. *International Studies Review*, 4(2), 123–135.

O'Cathain, A., Murphy, E., & Nicholl, J. (2008). Multidisciplinary, Interdisciplinary, or Dysfunctional? Team Working in Mixed-Methods Research. *Qualitative Health Research*, 18(11), 1574–1585.

O'Loughlin, K. (2001). *The Equivalence of Direct and Semi-Direct Speaking Tests*. Cambridge: Cambridge University Press.

O'Loughlin, K. (2011). The Interpretation and Use of Proficiency Test Scores in University Selection: How Valid and Ethical Are They? *Language Assessment Quarterly*, 8(2), 146–160.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing*, 30(3), 363–380.

O'Sullivan, B. (2016). A Story to Tell, a Lesson to Learn: The Testing Industry and Validation. Presented at the *ALTE 48th Conference*. Stockholm.

O'Sullivan, B., & Weir, C (2011). Testing and Validation, in B. O'Sullivan (ed.) *Language Testing: Theory and Practice* (pp. 13–32). Oxford: Palgrave.

O'Toole, L. (2000). Research on Policy Implementation: Assessment and Prospects. *Journal of Public Administration Research & Theory*, 10(2), 263–288.

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

Onderwijs Vlaanderen. (2015). *Curriculum: Eindtermen, ontwikkelingsdoelen, basiscompetenties en doelen beroepsgerichte vorming*. Retrieved April 11, 2016, from www.ond.vlaanderen.be.

Ongenaert, D. (2015). Buitenlandse studenten in Vlaanderen: wie zijn ze en wat doen ze? Retrieved September 22, 2015, from http://mo.be.

Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics*, 24(4), 492–518.

Ortega, L. (2008). *Understanding Second Language Acquisition*. London: Routledge.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, 27(2), 261–282.

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and Validating Band Levels and Descriptors for Reporting Overall Examinee Performance. *Language Assessment Quarterly*, 12(2), 153–177.

Paradis. (2007). Early bilingual and multilingual acquisition. In P. Auer & L. Wei (Eds.), *Handbook of Multilingualism and Multilingual Communication* (pp. 15–45). New York: Mouton de Gruyter.

Patton, M. (2002). *Qualitative Evaluation and Research Methods*. Newbury Park, CA: Sage.

Pellegrino Aveni, V. A. (2005). *Study Abroad and Second Language Use: Constructing the Self*. Cambridge: Cambridge University Press.

Pérez Vidal, C., & Juan-Garau, M. (2014). The effect of context and input conditions on oral and written development: A Study Abroad perspective. In C. Pérez-Vidal (Ed.), *Language Acquisition in Study Abroad and Formal Instruction Contexts* (pp. 157–185). Amsterdam: John Benjamins Publishing Company.

Pérez-Tattam, Mueller Gathercole, Yavaz, & Stadthagen-González. (2013). Measuring grammatical knowledge and abilities in bilinguals: implications for assessment and testing. In V. C. M. Mueller Gathercole (Ed.), *Issues in the Assessment of Bilinguals* (pp. 111–129). Bristol: Multilingual Matters.

Pérez-Vidal, C. (2014). *Language Acquisition in Study Abroad and Formal Instruction Contexts*. Amsterdam: John Benjamins Publishing Company.

Peters, E., & Van Houtven, T. (2010). *Taalbeleid in het hoger onderwijs: de hype voorbij?* Leuven: Acco.

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.

Polio, C. (2013). The acquisition of second language writing. In S. M. Gass & A. Mackey (Eds.), *The Routledge Handbook of Second Language Acquisition* (pp. 319–334). London; New York: Routledge.

Porto, M. (2012). Academic perspectives from Argentina. In M. Byram and L. Parmenter (Eds.), *The Common European Framework of Reference: The Globalisation of Language Education Policy* (pp. 129–38). Bristol: Multilingual Matters.

Purpura, J., E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied language research. *Language Learning*, 65(1), 36-73.

Ranta, L., & Meckelborg, A. (2013). How Much Exposure to English Do International Graduate Students Really Get? Measuring Language Use in a Naturalistic Setting. *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, 69(1), 1–33.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rawls, J. (2001). *Justice as Fairness: A Restatement*. Cambridge, MA: 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.

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.

Rijlaarsdam, Braaksma, Couzijn, Janssen, Kieft, & van den Bergh. (2005). Psychology and the teaching of writing in 8000 and some words. *British Journal of Educational Psychology Monograph Series II: Psychological Aspects of Education - Current Trends*, 3, 127–153.

Roever, C., & McNamara, T. (2006). Language testing: the social dimension. *International Journal of Applied Linguistics*, 16(2), 242–258.

Ross, S. J. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing*, 25(1), 5–13.

Salamoura, A. & Saville, N. (2009). Criterial features across the CEFR levels: Evidence from the English Profile Programme. *Research Notes*, 37, 34–40.

Sanz, C. (2014). Contributions of study abroad research to our understanding of SLA processes and outcomes: The SALA Project, an appraisal. In C. Pérez Vidal (Ed.), *Language Acquisition in Study Abroad and Formal Instruction Contexts* (pp. 1–17). Amsterdam: John Benjamins Publishing.

Sasayama, S. (2016). Is a "Complex" Task Really Complex? Validating the Assumption of Cognitive Task Complexity. *The Modern Language Journal*, 100(1), 231–254.

Schoonen, R., Gelderen, A. van, Glopper, K. de, Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First Language and Second Language Writing: The Role of Linguistic Knowledge, Speed of Processing, and Metacognitive Knowledge. *Language Learning*, 53(1), 165–202.

Schwartz, N., Knäuper, B., Oyersman, D., & Stich, C. (2008). The psychology of asking questions. In E. de Leeuw, J. Hox, & D. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 18 – 35). New York: Routledge.

Seloni, L. (2012). Academic literacy socialization of first year doctoral students in US: A micro-ethnographic perspective. *English for Specific Purposes*, 31(1), 47–59.

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.

Shaw, S., & Imam, H. (2013). Assessment of International Students Through the Medium of English: Ensuring Validity and Fairness in Content-Based Examinations. *Language Assessment Quarterly*, 10(4), 452–475.

Shohamy, E. (1994). The Validity of Direct Versus Semi-direct Oral Tests. *Language Testing*, 11, 99–123.

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

Shohamy, E. (2006). *Language Policy: Hidden agendas and new approaches*. London & New York: Routledge.

Shohamy, E. (2011). An engagement with the CEFR. A critical view and time. Paper presented at the *ALTE 4th International Conference*, Krakow, Poland.

Sin, C. (2003). Interviewing in "place": the socio-spatial construction of interview data. *Area*, 35(3), 305–312.

Snow, C. (2010). Academic Language and the Challenge of Reading for Learning About Science. *Science*, 328(5977), 450–452.

Song, M.-Y. (2012). Note-taking quality and performance on an L2 academic listening test. *Language Testing*, 29(1), 67–89.

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

Spolsky, B. (2004). *Language Policy*. Cambridge: Cambridge University Press.

Spolsky, B. (2008). Introduction: Language Testing at 25: Maturity and responsibility? *Language Testing*, 25(3), 297–305.

Stæhr, L. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139–152.

Steemans, S., & Vlasselaers, A. (2013). B2 or Not B2? Mapping the Oral Part of the ITNA to the CEFR. In J. Colpaert, M. Simons, A. Aerts, & M. Oberhofer (Eds.), *Language Testing in Europe: Time for a New Framework?* (pp. 101–102). Antwerp: University of Antwerp.

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing*, 18(2), 103–118.

Stricker, L. J. (2004). The performance of native speakers of English and ESL speakers on the computer-based TOEFL and GRE General Test. *Language Testing*, 21(2), 146–173.

Strobbe, L. (2016). Taalbeleid of talenbeleid? De plaats van meertaligheid op school. In L. Van Praag, S. Sierens, O. Agirdag, P. Lambert, S. Slembrouck, P. Van Avermaet, … M. Van Houtte (Eds.), *Haal meer uit meertaligheid. Omgaan met talige diversiteit in het basisonderwijs* (pp. 117–130). Leuven: Acco.

Subtirelu, N. (2014). A language ideological perspective on willingness to communicate. *System*, 42, 120–132.

Swain, M., & Deters, P. (2007). "New" Mainstream SLA Theory: Expanded and Enriched. *The Modern Language Journal*, 91, 820–836.

Swender, E. (2010). A Tale of Two Tests. STANAG and CEFR. Comparing the Results of side-by-side testing of reading proficiency. Paper presented at *BILC*, Istanbul.

Tannenbaum, R.J., & Wylie, C.E. (2008). *Linking English-Language Test Scores onto the Common European Framework of Reference: An Application of Standard-Setting Methodology*. Princeton, NJ: ETS.

Taylor, C. (1992). The Politics of recognition. In A. Gutmann (Ed.), *Multiculturalism and "the politics of recognition"* (pp. 25–75). Princeton, N.J: Princeton University Press.

Taylor, L. (2004). Issues of test comparability. *Cambridge Research Notes*, 15, 2-5.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21–36.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing*, 30(3), 403–412.

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.

The Douglas Fir Group. (2016). A Transdisciplinary Framework for SLA in a Multilingual World. *The Modern Language Journal*, 100(S1), 19–47.

Tillema, M., Bergh, H. van den, Rijlaarsdam, G., & Sanders, T. (2013). Quantifying the quality difference between L1 and L2 essays: A rating procedure with bilingual raters and L1 and L2 benchmark essays. *Language Testing*, 30(1), 71–97.

Toulmin, S. (2001). *Return to Reason*. Cambridge, Mass.: Harvard University Press.

Toulmin, S. (2003). *The Uses of Argument*. Cambridge, U.K. ; New York: Cambridge University Press.

Truyts, J., & Torfs, M. (2015, January 22). Bij Nederlandse overrompeling wordt het ingewikkeld. Retrieved September 22, 2016, from http://deredactie.be.

Tschirner, E., Bärenfänger, O., & Wisniewski, K. (2015). *Assessing Evidence of Validity of the ACTFL CEFR Listening and Reading Proficiency Tests (LPT and RPT) Using a Standard-Setting Approach. (Technical Report 2015-EU-PUB-2)*. Leipzig: Institute for Test Research and Test Development.

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

UN General Assembly. (1948). *Universal Declaration of Human Rights*. UN General Assembly. Retrieved from http://www.ohchr.org.

Universiteit Antwerpen. (2016). *PROCEDURE PROC/ADOND/001.1*. Retrieved on April 12, 2016, from www.uantwerpen.be.

Universiteit Gent. (2013). *Registratie van kansengroepen aan de UGent*. Retrieved on April 1, 2016, from www.ugent.be.

Universiteit Gent. (2015). *Onderwijs- en examenreglement 2015-2016*. Retrieved on 12 January, 2015, from www.ugent.be.

Universiteit Gent. (2016). *Education and Examination Code academic year 2016-2017.* Retrieved September 30, 2016, from www.ugent.be.

Universiteit Hasselt. (2015) *Toelatingsvoorwaarden*. Retrieved on 29 january, 2015, from www.uhasselt.be.

Universiteit Hasselt. (2016). Taalvoorwaarden. Retrieved on 29 january, 2015, from www.uhasselt.be.

University of Cambridge, ESOL Examinations. (2011). *Using the CEFR: Principles of Good Practice*. Cambridge: University of Cambridge, ESOL Examinations.

Van Avermaet, P. & Pulinx, R. (2013). Language testing for immigration to Europe. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 376–389). Malden, MA, USA: John Wiley & Sons, Inc.

Van Avermaet, P., & Gysen, S. (2006). From needs to tasks: Language learning needs in a task-based approach. In K. van den Branden (Ed.), *Task-Based Language Education: From Theory to Practice* (pp. 17–46). Cambridge: Cambridge University Press.

Van Avermaet, P., & Slembrouck, S. (2014). In een taalbad verzuip je. Retrieved on May 17, 2016, from www.standaard.be.

Van den Bosch, K., & Cantillon, B. (2006). Policy impact. In M. Moran, M. Rein, & R. E. Goodin (Eds.), *The Oxford handbook of public policy* (pp. 296–319). Oxford: Oxford University Press.

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

Van Ek, J. & Trim, J. (1991a). *Threshold 1990*: Strasbourg : Council of Europe.

Van Ek, J. & Trim, J. (1991b). *Waystage 1990*. Strasbourg: Council of Europe.

Van Ek, J. & Trim, J. (2001). *Vantage*. Cambridge: Cambridge University Press.

Van Gorp, K., Luyten, L., De Wachter, L., & Steemans, S. (2014). A concurrent validity study of two academic placement tests. Presented at the *ALTE 5th International Conference*, Paris.

Van Houtven Tine, Peters Elke (2010). De weg naar materiaalontwikkeling is geplaveid met behoeftes. In: Peters E., Van Houtven T. (Eds.), Taalbeleid in het hoger onderwijs: de hype voorbij?, (pp. 69-85). Leuven: Acco.

Van Houtven, T., Peters, E., & El Morabit, Z. (2010). Hoe staat het met de taal van studenten? Exploratieve studie naar begrijpend lezen en samenvatten bij instromende studenten in het Vlaamse hoger onderwijs. *Levende Talen Tijdschrift*, 11(3), 29–44.

Van Splunder, F. (2015). *Taalstrijd. Over relaties tussen talen in de wereld, Europa, Nederland en Vlaanderen*. Brussels: ASP.

Van Weijen, D. (2009). *Writing processes, text quality, and task effects: Empirical studies in first and second language writing*. Utrecht: LOT Dissertation Series.

Van Weijen, D., Van den Bergh, B., Rijlaarsdam, G., & Sanders, T. (2008). Differences in process and process-product relations in L2 writing. *ITL International Journal of Applied Linguistics*, 156, 203–226.

Vanbelle, S., & Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*, 6(2), 157–163.

Vedung, E. (2013). Six models of evaluation. In E. J. Aral, S. Fritzen, M. Howlett, M. Ramesh, & X. Wu (Eds.), *Routledge handbook of public policy* (pp. 387–400). London & New York: Routledge.

Victori, M. (1999). An analysis of writing knowledge in EFL composing: a case study of two effective and two less effective writers. *System*, 27(4), 537–555.

Vlaamse Regering. (2013). *Besluit van de Vlaamse Regering tot codificatie van de decretale bepalingen betreffende het hoger onderwijs, Pub. L. No. B.S.27/02/2014*, § Deel 2. Hoofdstuk 8. Retrieved on December 11, 2016, from http://data-onderwijs.vlaanderen.be.

Vrije Universiteit Brussel. (2014). *Onderwijs- en examenreglement 2014-2015*. Retrieved on October 25, 2016, from http://www.vub.ac.be.

Wainer, H. (2011). *Uneducated guesses: Using evidence to uncover misguided education policies*. Princeton, N.J: Princeton University Press.

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

Walton, G. M., & Cohen, G. L. (2007). A Question of Belonging: Race, Social Fit, and Achievement. *Journal of Personality and Social Psychology*, 92(1), 82–96.

Wang, S., Wang, N., & Hoadley, D. (2007). Construct Equivalence of a National Certification Examination that Uses Dual Languages and Audio Assistance. *International Journal of Testing*, 7: 255–68.

Wartenbergh, F., Brukx, D., van den Broek, A., Jacobs, J., Pass, J., Hogeling, L., & van Klingeren, M. (2009). *Studentenmonitor Vlaanderen 2009*. Departement Onderwijs en Vorming.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Weigle, S. C., & Friginal, E. (2015). Linguistic dimensions of impromptu test essays compared with successful student disciplinary writing: Effects of language background, topic, and L2 proficiency. *Journal of English for Academic Purposes*, 18, 25–39.

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). London and New York: Routledge.

Weir, C. (2005a). *Language Testing and Validation*. New York: Palgrave.

Weir, C. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22(3), 281–300.

Welkenhuysen-Gybels, J., & van de Vijver, F. (2001). A Comparison of Methods for the Evaluation of Construct Equivalence in a Multigroup Setting. In *Proceedings American Statistical Association* (CD-ROM), Atlanta.

Welsh, E. (2002). Dealing with Data: Using NVivo in the Qualitative Data Analysis Process. *Forum: Qualitative Social Research*, 3(2). Retrieved on July 30, 2015, from http://www.qualitative-research.net.

Wet op het hoger onderwijs en wetenschappelijk onderzoek. (1992). Pub. L. No. BWBR0005682, § 7.2.

Whitehead, A. N. (1925). *Science and the modern world*. New York: Pelican Books.

Wilson, R. (2006). Policy analysis as policy advice. In M. Moran, M. Rein, & R. E. Goodin (Eds.), *The Oxford handbook of public policy* (pp. 152–169). Oxford: Oxford University Press.

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing*, 27, 1–10.

Wollmann, H. (2007). Policy evaluation and evaluation research. In F. Fischer & G. J. Miller (Eds.), *Handbook of Public Policy Analysis: Theory, Politics, and Methods* (pp. 393–402). Boca Raton: CRC Press.

Wu, R.-J. (2013). Native and non-native students' interaction with a text-based prompt. *Assessing Writing*, 18(3), 202–217.

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

Xi, X., Bridgeman, B. & Wendler, C. (2014). Tests of English for Academic Purposes in University Admissions. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 318–33). Malden, MA: John Wiley & Sons, Inc.

Yu, G. (2013a). From Integrative to Integrated Language Assessment: Are We There Yet? *Language Assessment Quarterly*, 10(1), 110–114.

Yu, G. (2013b). The Use of Summarization Tasks: Some Lexical and Conceptual Analyses. *Language Assessment Quarterly*, 10(1), 96–109.

Zheng, Y, & De Jong, J.H.A.L. (2011). *Research Note: Establishing Construct and Concurrent Validity of Pearson Test of English Academic*. Retrieved on September 29, 2016, from http://pearsonpte.com.

ACADEMIC OUTPUT RELATED TO THIS PHD

Peer reviewed journal articles

Deygers, B. (2017, accepted). A year of highs and lows. Considering contextual factors to explain L2 gains at university. *The Modern Language Journal.*

Deygers, B., Van den Branden, K., & Peters, E. (2017). Checking assumed proficiency: Comparing L1 and L2 performance on a university entrance test. *Assessing Writing*, 32, 43–56.

Deygers, B., Van den Branden, K., Van Gorp, K. (2017, in press). University entrance language tests: a matter of justice. *Language Testing.*

Deygers, B., Van Gorp, K., & Demeester, T. (2017, in press). The B2 level and the dream of a common standard. *Language Assessment Quarterly*.

Deygers, B., Zeidler, D., Vilcu, D., & Carlsen C.H. (2017, in press). One framework to unite them all? The use of the CEFR in European university entrance policies. *Language Assessment Quarterly*.

Deygers, B. & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, 32(4), 521–541.

Peer reviewed book chapters

Deygers, B. (2017, in press). University entrance language tests: examining assumed equivalence. In J. Davis, J. Norris, M. Malone, T. McKay, & Y Son (Eds.). *Useful Assessment And Evaluation In Language Education.* Washington, D.C.: Georgetown University Press.

Presentations

Plenary

Deygers, B. (2016). Na de toets. De ervaringen van anderstalige studenten aan Vlaamse universiteiten. *Forumdag Taalbeleid Hoger Onderwijs*. Universiteit Antwerpen.

Deygers, B. (December, 2015). After the entrance test: A longitudinal study of L2 students' experiences at university. *Advanced Research in Multilingualism Talk Series #9*. Georgetown University, Washington, D.C.

Deygers, B. (2015). Language tests for university admission: a mismatch between test and practice? *EALTA*, Copenhagen.

Deygers, B. & Zeidler, B. (2015). The CEFR and university entrance tests – a state of affairs in Europe. *ALTE*, Bergen.

Deygers, B. & Van Gorp, K. (2014). Keeping it Real: Task-Based language assessment. *ALTE*, London.

Deygers, B. (2014). Creating meaning out of vagueness? Developing a rating scale from general descriptors. *NATO/BILC Conference*, Bruges. Invited plenary.

Parallel (selection)

Deygers, B. (2016). A longitudinal study of L2 students' experiences at university. *LTRC*, Palermo.

Deygers, B. (2016). University entrance language tests: Justice at the core of the construct. *GURT*. Georgetown University, Washington, D.C.

Deygers, B. (2015). Concurrent validity in university entrance tests: discrepancies below the surface correlation. *ECOLT*, Washington, D.C.

Deygers, B. (2015). Justice and validity in university entrance test tasks: The case of Flanders. *TBLT*, Leuven.

Deygers, B. (2015). The concurrent and predictive validity of university entrance tests. *LTRC*, Toronto.

Deygers, B. & Van Gorp; K. (2014). Creating meaning out of vagueness? Using the CEFR as the foundation for an academic rating scale. LTRC, Amsterdam.

Deygers, B. & Carlsen, C.H. (2014). The B2 level and its applicability in university Placement tests. *ALTE* 5th International Conference, Paris.

Deygers, B. & Van Gorp, K. (2013). Examining the concurrent validity of a task-based and an indirect academic language test. *TBLT*. Banff.

SAMENVATTING

Internationale L2 studenten kunnen zich pas inschrijven aan een Vlaamse universiteit wanneer ze voldoen aan de talige eisen. Aan elke universiteit is het basisniveau dat van deze studenten verwacht wordt B2 op het ERK (Common European Framework of Reference for Languages – Council of Europe, 2001). Studenten kunnen op verschillende manieren bewijzen dat ze voldoen aan de taaleisen. Ze kunnen een taaltoets afleggen, maar ze kunnen ook dadelijk starten wanneer ze minstens één jaar in het Nederlandstalige secundair of hoger onderwijs succesvol hebben afgerond.

De twee B2 tests die aan elke universiteit aanvaard worden, zijn ITNA en STRT. Deze tests hebben hetzelfde doel en hetzelfde ERK niveau, maar kennen enkele substantiële verschillen in operationalisering. ITNA heeft een computergestuurde schriftelijke component en bestaat uit gesloten vraagtypes die vooral beroep doen op receptieve vaardigheden en woordenschat- en grammaticakennis. De schriftelijke component van STRT daarentegen is taakgericht en geïntegreerd. De mondelinge secties van STRT en ITNA lijken wel sterk op elkaar; ze bestaan uit een presentatie- en een argumentatietaak, en hebben vijf overlappende beoordelingscriteria die gebaseerd zijn op dezelfde ERK-descriptoren.

Dit onderzoeksproject onderzocht de voornaamste assumpties die aan de basis liggen van het toelatingsbeleid, vanuit drie perspectieven om na te gaan hoe effectief het toelatingsbeleid van Vlaamse universiteiten is ten aanzien van internationale L2 studenten: constructen en niveaus (meten de toetsen de zaken die belangrijk zijn op het juiste niveau), selectie (selecteren de toetsen dezelfde kandidaten), en taalevolutie na de toets.

Constructen & niveaus

De eerste studie onderzocht het toelatingsbeleid voor internationale L2 studenten universiteiten in 28 Europese regio's. Uit deze bevraging bleek dat B2 veruit het meest gevraagde toelatingsniveau is, en dat in de meeste regio's verschillende tests naast elkaar aanvaard worden. In die zin is het Vlaamse beleid representatief voor wat er doorgaans binnen Europa gebeurt.

Hoofdstuk 2 vergeleek de operationalisering van STRT en ITNA met de reële talige eisen aan de universiteit, en onderzocht in hoeverre slagen voor STRT en ITNA ook impliceert dat men klaar is voor de talige uitdagen die volgen. De studie combineerde de meningen en ervaringen van 24 universitaire medewerkers en 31 internationale L2 studenten, van wie er twintig longitudinaal werden gevolgd nadat ze STRT en ITNA afgelegd hadden. Uit de resultaten bleek dat de werkelijke taaleisen aan de Vlaamse universiteiten soms cruciaal afwijken

van de inhoud van beide tests; dat L2 studenten die geslaagd waren voor ITNA of STRT, of beide, lang niet klaar waren voor de receptieve eisen van de academische wereld; en dat vier van de zeven studenten die niet geslaagd waren voor STRT of ITNA toch goed presteerden aan de universiteit. Uit deze studie bleek tevens dat het B2 niveau als toelatingsniveau onvoldoende garanties biedt dat instromende internationale L2 studenten zullen voldoen aan de talige eisen van de universiteit.

Selectie & equivalentie

In twee studies werd nagegaan of STRT en ITNA equivalente B2 toetsen kunnen zijn. De eerste studie was gericht op equivalentie qua niveau en qua construct, en het tweede onderzoek richtte zich specifiek op de gelijkwaardigheid van overeenkomstige ERK-gebaseerde criteria. Op basis van de scores van 118 deelnemers die STRT en ITNA binnen dezelfde week aflegden, toonde de eerste studie aan dat de totale correlatie tussen STRT en ITNA scores matig hoog was (r = 0,767 **), net zoals de correlatie tussen de schriftelijke onderdelen (r = 0,694 **). De overeenkomst tussen de scores op het mondeling examen was echter veel lager (τ = 0,387 **). Uit de aanvullende analyses kwamen verdere discrepanties naar voren. Allereest bleek de kans om te slagen voor STRT (50%) significant (p = 0,02) groter dan de ITNA-slaagkans (35%). Ten tweede toonden lineaire regressie en Rasch analyses aan dat er belangrijke verschillen bestaan tussen de constructs van STRT en ITNA. De woordenschat- en grammaticataken van ITNA zijn moeilijker dan alle andere schriftelijke ITNA of STRT taken, terwijl de argumentatieve taken van STRT de makkelijkste geschreven taken bleken. Bovendien duidde de Rasch analyse betrouwbaar (0,88) aan dat de gesproken component van ITNA moeilijker is dan die van STRT. Ook hier was de reden voor de discrepantie de relatieve moeilijkheid van de taalkundige criteria bij ITNA versus de relatieve mildheid van inhoudelijke criteria bij STRT.

De tweede studie vergeleek scores op corresponderende criteria binnen de mondelinge onderdelen van STRT en ITNA. Deze componenten bevatten zeer vergelijkbare taaktypes en criteria: beide tests bevatten vijf criteria die gebaseerd zijn op dezelfde ERK descriptoren. De analyses (lineaire en meervoudige regressie en Rasch) toonden dat ITNA en STRT elk ERK-gebaseerd criterium anders interpreteerden. Voor alle corresponderende criteria waren de gewogen kappa coëfficiënten laag ($Kw \leq$ 0,216) en was de correlatie laag. Bovendien gaf het Rasch model betrouwbaar aan dat overeenkomstige criteria nooit dezelfde moeilijkheidsgraad hadden.

Hoofdstuk vijf verifieerde de veronderstelling dat Vlaamse studenten, die geen taaltest moeten afleggen om toelating te krijgen tot de universiteit, het B2 niveau de facto hebben bij instroom. Indien niet alle Vlaamse eerstejaarsstudenten dit niveau bereiken, slaagt het toelatingsbeleid er niet in om

het B2 minimumniveau onder de eerstejaarsstudenten te garanderen. Om dit te onderzoeken, legden 159 Vlaamse eerstejaarsstudenten twee schriftelijke STRT taken af tijdens de eerste maand van het universitair onderwijs. Met behulp van niet-parametrische statistiek en Rasch analyse werden de L1 scores vergeleken twee groepen van L2 kandidaten. De resultaten toonden aan dat L1 studenten over het algemeen hoger scoorden dan L2 kandidaten, maar ook dat L2 kandidaten hogere scores haalden op de inhoudelijke criteria. Dat Vlaamse kandidaten het beter deden op vlak van formele criteria en op vlak van algemene scores impliceert echter niet dat alle Vlaamse studenten slaagden: 11% van de Vlaamse studenten haalde het B2 niveau niet.

Taalevolutie na de toets

Tot op heden is er weinig onderzoek gedaan naar hoe internationale L2 studenten zich talig redden in de doelcontext tijdens de maanden na de toelatingsproef. Nog minder studies hebben dit punt onderzocht vanuit een kwalitatief en longitudinaal perspectief. In dit onderzoek werden 20 internationale L2 studenten gevolgd tijdens hun eerste jaar aan een Vlaamse universiteit. Tijdens die periode werden zij maandelijks geïnterviewd en legden ze na acht maanden opnieuw twee STRT taken af. De resultaten toonden aan dat de respondenten geen significante vooruitgang hadden gemaakt in termen van STRT-score, of in termen van courante maten van complexiteit, nauwkeurigheid, of vlotheid. Het enige significante verschil was een verminderde hoeveelheid woorden tijdens de mondelinge presenteeropdracht. Uit de analyses bleek dat bijna alle respondenten een sociaal, institutioneel en academisch isolement ervaren hadden. De internationale L2 studenten hadden dus slechts beperkte mogelijkheden gehad om te interageren met Vlaamse sprekers, wat zeer waarschijnlijk heeft bijgedragen tot de beperkte positieve taalevolutie.

Conclusie

In het vierde en laatste deel van dit onderzoek werd bekeken hoe het Vlaamse toelatingsbeleid tot stand komt. Beleidsmakers op Vlaams niveau en op niveau van de vijf universiteiten werden bevraagd. Uit de analyse van de interviews kwam naar voren dat het Vlaamse beleid niet in de eerste plaats stoelt op empirisch onderzoek, maar eerder resulteert uit de belangen van belangrijke stakeholders en uit het uitwerken van oplossingen voor ad-hoc problemen.

De resultaten van het onderzoek bieden slechts beperkte argumenten om de effectiviteit van het toelatingsbeleid te ondersteunen. Het lijkt onwaarschijnlijk dat het huidige toelatingsbeleid in staat is om een B2 taalniveau binnen de gehele studentenpopulatie te waarborgen, aangezien de gebruikte tests niet als

gelijkwaardig kunnen worden beschouwd, en aangezien meer dan een op tien Vlaamse studenten niet slaagde voor een schriftelijke B2 toets. Bovendien is gebleken dat het B2-niveau lager ligt dan de reële receptieve taaleisen aan de universiteit en dat de taken binnen de taaltoetsen niet steeds in overeenstemming zijn met de werkelijke taalopdrachten. Ten slotte maken internationale L2 studenten tijdens hun eerste jaar vermoedelijk weinig talige vooruitgang, en lijken de taaltoetsen weinig tot geen positief effect te hebben op de maatschappelijke integratie van internationale L2 studenten.

Appendix 1 (1/3). STRT Part 1: Listening-into-writing

	Task 1		Task 2	
Task goal	Write an argumentative text based on radio fragment		Listen to a lecture and to summarize it for fellow students	
Instruction	▪ Write a summary and an argument ▪ Total time: 29 minutes		▪ Write a summary ▪ Total time: 50 minutes	
Receptive demands	Four-minute scripted radio fragment ▪ Listen once ▪ Topic: using laptops versus pen and paper to take class notes ▪ Even pace (152 words/minute) ▪ Clear pronunciation		Nine-minute scripted lecture ▪ Listen twice ▪ Topic: industrialization ▪ Even pace (126 words/minute) ▪ Clear pronunciation	
Productive demands	Write 180+ words The performance should ▪ be understandable for people unfamiliar with the broadcast; ▪ contain an introduction, body, and a conclusion; ▪ provide four arguments.		Write 160+ words The performance should ▪ be helpful for people unfamiliar with the topic; ▪ include the main points of the lecture; ▪ follow a preset structure.	
Criteria	Content Vocabulary Grammar Cohesion Mechanics	9 binary criteria (e.g. sound arguments) CEFR-linked 4-band scale 4 ≥ C1 3 = B2 2 = B1 1 ≤ A2	Content Vocabulary Grammar Cohesion Mechanics	9 binary criteria (e.g. conditions for industrialization) CEFR-linked 4-band scale 4 ≥ C1 3 = B2 2 = B1 1 ≤ A2

Appendix 1 (2/3). STRT, Part 2: Reading-into-writing

	Task 3		Task 4	
Task goal	Write a formal letter to your university's examination committee		Read a popularizing paper, summarize it and formulate an argument.	
Instruction	▪ Read the description of four study programs and apply for two ▪ Total time: 45 minutes		▪ Summarize the article and explain own viewpoint ▪ Total time: 60 minutes	
Receptive demands	Four texts describing study programs ▪ 311 words combined ▪ Flesh Reading Ease: 30 ▪ Estimated grade level: 11		Article concerning gender differentiation in education ▪ 910 words ▪ Flesh Reading Ease: 49 ▪ Estimated grade level: 9	
Productive demands	Write 130+ words The performance should ▪ include a question for the committee; ▪ describe chosen study programs; ▪ include a reason for choosing each program.		Write 250+ words The performance should ▪ explain the issue discussed; ▪ mention the main research results; ▪ provide two arguments to substantiate one's opinion; ▪ include a conclusion.	
Criteria	Content Vocabulary Grammar Cohesion Mechanics Register	8 binary criteria (e.g. program description) CEFR-linked 4-band scale	Content Vocabulary Grammar Cohesion Mechanics Register	8 binary criteria (e.g. research results) CEFR-linked 4-band scale

4 ≥ C1
3 = B2
2 = B1
1 ≤ A2

Appendix 1 (3/3). STRT, Part 3: Speaking

	Task 5		Task 6	
Task goal	Motivate a chosen internship position		Give a presentation about ocean pollution	
Instruction	▪ Read information about two internship positions ▪ Choose one, and motivate choice ▪ Total time: 10 minutes (5 minutes preparation)		▪ Go through the slides, read the background article ▪ Give a presentation and answer the questions ▪ Total time: 15 minutes (10 minutes preparation)	
Receptive demands	Brief written description of internship positions o table layout, bullet points o 187 words		▪ 7 slides with bullet points, graphs and tables ▪ Popularizing article o 305 words o Flesh Reading Ease: 49 o Grade level: 10	
Productive demands	Five-minute conversation The performance should ▪ include three arguments; ▪ provide adequate answers to the examiner's questions.		Five-minute presentation The performance should ▪ discuss all slides; ▪ explain a graph; ▪ include possible solutions to the problem presented; ▪ provide adequate answers to the examiner's questions.	
Criteria	Content Vocabulary Grammar Register Pronunciation Fluency Initiative	9 binary criteria (e.g. conditions for industrialization) CEFR-linked 4-band scale	Content Vocabulary Grammar Cohesion Pronunciation Fluency Initiative	15 binary criteria (e.g. information about every slide) CEFR-linked 4-band scale
		4 ≥ C1 3 = B2 2 = B1 1 ≤ A2		4 ≥ C1 3 = B2 2 = B1 1 ≤ A2

Appendix 2 (1/2). ITNA: computer test

	Format	Operationalization
Language-in-use		
Vocabulary	Multiple choice	Pick one of four options to complete a sentence
	Multiple choice	Select one of four expressions which best fits a given description
	Short open answer	Word transformation so entry matches a given sentence
Grammar	Short open answer	Word transformation so entry matches a given sentence
Cloze	Short cloze texts	Three texts with five gaps each (selected words are omitted, rather than every n^{th} word): popularizing scientific text
	Long cloze text	One text with ten gaps (selected words are omitted, rather than every n^{th} word): popularizing scientific text
Reading		
Comprehension	Multiple choice	Five short newspaper articles (around 300 words), with two comprehension questions each
Structure	Drag-and-drop	First and last sentence are given, candidates restructure five jumbled sentences
Listening		
Comprehension	Multiple choice	Three four-minute radio extracts, with four comprehension questions each
Dictation	Short open answer / dictation	Write down eight words as mentioned in natural speech (news broadcast)

Appendix 2 (2/2). ITNA: Speaking test

	Argumentation	Presentation
Task goal	Formulate an argument	Give a presentation about a general topic (i.e., smoking)
Instruction	▪ Read information about two possible topics ▪ Choose one topic, and formulate an argument ▪ Total time: 10 minutes (15 minutes preparation for both tasks)	▪ Go through the slides ▪ Give a presentation and answer the questions ▪ Total time: 5 minutes
Receptive demands	Brief written description of topics	▪ 5 slides with bullet points, graphs and/or tables
Productive demands	Five-minute conversation The performance should ▪ include three arguments; ▪ provide adequate answers to the examiner's questions.	Three-minute presentation The performance should ▪ discuss all slides; ▪ explain two graphs and/or tables; ▪ provide adequate answers to the examiner's questions.
Criteria	Vocabulary Grammar Pronunciation Fluency Cohesion	CEFR-linked 4-band scale 4 ≥ C1 3 = B2 2 = B1 1 ≤ A2

Appendix 3. $L2_P$ participants

	M/F[*]	**L1**	**Nationality**	**U**[⋆]	**B/M**[°]	**Faculty**	**L2**[♯]	**Test**[+]
Heddi	F	German	Switzerland	G	B	Engineering	24	STRT
Noor	F	Farsi	Iran	G	M	Medicine	17	ITNA
Hassan	M	Turkish	Turkey	G	M	Law	9	ITNA
Vincent	M	French	France	G	M	Sciences (Geology)	12	ITNA
Marion	F	Greek	Greece	G	M	Political sciences	48	ITNA
Sandrine	F	French	Belgium	G	M	Law	12	ITNA
David	M	English	South Africa	G	B	Social sciences	36	ITNA
Henna	F	Finnish	Finland	G	M	Psychology	14	ITNA
Abdullah	M	Arabic	Morocco	G	M	Sciences (Chemistry)	12	ITNA
Hatice	F	Turkish	Turkey	G	M	Political sciences	36	ITNA
Janet	F	English	Cameroon	G	B	Medicine	12	ITNA

Note. [*]Male/Female
[°]Bachelor/Master
[♯]months of Dutch L2 instruction
[+]university entrance test taken

Appendix 4. $L2_F$ participants

	M/F*	Age	L1	Nationality	U★	B/M°	Faculty	L2#	STRT	ITNA	+/-+
Elena	F	23	Ukrainian	Ukraine	G	M	Engineering	18	1	1	+
Alexandra	F	24	Spanish	Peru	G	M	Economics	7	1	1	+
Marie	F	19	French	Belgium	G	M	Law	120	1	1	+
Leila	F	44	Haitian	Haiti	I	M	Political sciences	20	1	0	+
Guadalupe	F	30	Spanish	El Salvador	G	B	Psychology	12	1	0	+
Elif	F	24	Turkish	Turkey	L	M	History	12	1	1	+
Oksana	F	21	Ukrainian	Ukraine	L	B	Linguistics	9	1	1	+
Ersi	F	19	Albanian	Albania	A	B	Law	8	1	1	+
Alireza	M	24	Farsi	Iran	G	B	Engineering	6	0	1	-
Merveille	F	27	French	Congo	L	M	Biomedical	10	1	0	-
Gabriela	F	23	Spanish	Costa Rica	L	B	Linguistics	11	1	0	-
Emma	F	21	German	Germany	L	B	Psychology	10	1	1	-
Hoang	M	24	Vietnamese	Germany/Vietnam	L	B	Psychology	12	1	1	-
Anastasia	F	26	Russian	Russia	L	M	Engineering	22	1	1	-
Océane	F	20	French	Belgium	A	M	Law	72	1	1	-
Clara	F	21	French	Belgium	A	M	Law	72	1	1	-
Stella	F	32	Armenian	Armenia	H	B	Economics	7	0	0	+/V
Yazdan	M	24	Pashto	Afghanistan	A	B	Economics	14	1	0	V
Jessica	F	27	Spanish	Chile	G	M	Medicine	24	1	0	?
Chloé	F	21	French	Belgium	L	B	Economics	72	1	1	?

Note. * Male/Female

★ Antwerp University /Ghent University / Interuniversity / University College of Hasselt / University of Leuven

° Bachelor/Master

months of Dutch L2 instruction

+ more(+)/less(-) than 50% of courses passed in July 2015/ Visa problems / reason for attrition unknown (?)

Appendix 5. 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 & Political Sciences	Professor	Ac4
	Tutor	Ac24