Radboud Universiteit

The effect of bilingual secondary education on the acquisition of vocabulary and non-native phonemic contrasts

Master's Thesis

RMA Linguistics & Communication Sciences

Lisanne Versaevel

S1065200

Nijmegen, 31 March 2023

First reader: Dr Eva Koch

Second reader: Prof Roeland van Hout

Acknowledgments

I am beyond glad to finally hand in this thesis and mark the end of my journey as a student as I do so. I began in Leiden as a student of English Language & Culture and straight away knew I loved linguistics best of all, and it's led me all the way to the Linguistics & Communication Sciences Research Master at Radboud to doing this thesis. In my six years as a student, I have learnt so much: not just about language, but also about how the world functions and my role in it. That is part of growing up, and your university and all the experiences during those years are beyond valuable; I don't think I would have been the person I am today without where all my choices have led me.

I would not be finishing my degree today without the help of several of these people I've met along the road. Firstly, I owe so much gratitude to Eva Koch, who was so willing to help me figure out my thesis project from the moment I met her. Your advice and help have been so valuable, Eva, as well as your mental support. I also want to thank Roeland van Hout who was never afraid to make sure to push me (mostly when it came to statistics!) and without whom I wouldn't have had this dataset to analyse.

Mostly, I want to thank my parents for supporting me in all my endeavours. Mum, thank you for always listening, because you knew I just needed to talk about what I was doing sometimes. Dad, I've loved all our conversations about language and hope we will have many more in the future. I also want to thank all the friends who have cheered me on throughout the years: you have done more for me than you know. I want to thank Vasudha especially, who made Nijmegen a home for me in the years we were both here.

And to Thijs, my support in all things: thank you for always being there.

Abstract

This thesis investigated how *Content Language Integrated Learning* (CLIL, which in Dutch is called TTO for "*tweetalig onderwijs*" (bilingual education)) in Dutch secondary education affected pupils' acquisition of English vocabulary and phonemic contrasts in English that are considered difficult for native Dutch speakers. Pupils were aged 12-16 and were all enrolled at the same high school, following either the monolingual track (Dutch) or the TTO track (Dutch and English). Two components of the same lexical decision task measured vocabulary size and phonemic discrimination ability, respectively. The performance on these two components was measured by means of both accuracy scores as well as d-prime scores in order to do a systematic comparison between both scores, allowing to assess which method of scoring should be preferred. An advantage of both TTO and school year was found for vocabulary size, with pupils who followed TTO performing better than non-TTO pupils and pupils in later school years outperforming pupils in lower years. No interaction effect was found for TTO and school year, indicating that TTO pupils already started out with a larger vocabulary size at the beginning of high school and maintained this lead, but did not improve more strongly over time than non-TTO pupils. Phonemic discrimination ability was measured with the /æ/-/ɛ/ vowel contrast and word-final voiced and unvoiced contrasts between fricatives and plosives. Remarkably, no main effects of TTO and school year were found for the pupils' scores on phonemic discrimination ability, which indicates that TTO and non-TTO pupils scored similarly on phonemic discrimination of the tested items and that pupils did not improve over time. Overall, pupils performed best on words with plosives and scored similarly on the discrimination between vowels and fricatives, but generally, most pupils' scores were low regardless of manner of articulation. A possible interpretation of the data is that CLIL does not offer enough exposure for pupils to implicitly learn to discriminate between difficult phonemic contrasts, or perhaps the quality of exposure is not sufficient. An additional possibility may be that, after a certain age, it is hard to improve one's phonemic discrimination ability.

Keywords: bilingual education, CLIL, phonemic discrimination, vocabulary size, second language acquisition, d-prime

Contents

1. Introduction

The use of English has become more and more prevalent in the Netherlands in recent decades, even causing researchers to ask whether English can still rightly be called a foreign language rather than a second language in Dutch society (Gerritsen et al., 2016). English has become strongly entrenched in everyday life, and perhaps it should not be surprising that English education is becoming more popular as well. In the Netherlands, the number of schools offering bilingual education has been increasingly growing in recent years (Van ’t Erve, 2021). The method of education these schools adhere to is known as *Content Language Integrated Learning.*

Content Language Integrated Learning (CLIL) is a form of bilingual education inspired by immigration programs, and seeks to immerse children in a new language. The aim is to teach new content in the target language, rather than having pupils acquire linguistic skills only during language classes (European Commission, 2006). With this approach, children acquire a new language in a communicative way, which should allow them to reach higher levels of proficiency (De Graaff, 2013). Research has focused on the role of bilingual education for pupils’ vocabulary size, grammatical skills, speaking ability, and reading skills among other things. Generally, CLIL has been found to have a positive influence on pupils’ language abilities (Admiraal et al., 2006; Huibregtse et al., 2002; Lorenzo et al., 2010). However, doubts remain about whether pupils improve indeed as strongly as they are often said to do (Bruton, 2011) or about whether pupils truly have a larger vocabulary size or better writing skills than those who do not follow CLIL (Agustín Llach, 2017). Moreover, hardly any research has been conducted into the effects of CLIL on phonemic discrimination, which is an important area of language acquisition that determines our ability to correctly comprehend and produce speech. Lastly, there is not one standard way to implement CLIL, and schools often vary in how it is applied (depending, for instance, on the time allotted for a foreign language, or the skills and availabilities of teachers) with some schools being more successful than others in teaching the target language (Olsson, 2021). This can cause contradictory findings and uncertainty about the influence of CLIL on the various areas of language acquisition that studies have focused on.

This thesis aims to address the gap on research about the role of CLIL on discriminating English phonemic contrasts that are considered difficult for native speakers of Dutch. A study was conducted in a high school in the Netherlands offering bilingual education. The same school also allows pupils to follow their education fully in Dutch which allows for a comparison between the bilingual and monolingual tracks. Vocabulary knowledge and phonemic discrimination ability were both measured using one lexical decision task. The phonemic contrasts in English were chosen that are known to be difficult for native speakers of Dutch so that we can assess the improvement on these contrasts over time. The combination of phonemic discrimination ability and vocabulary size in one task allows to gain a deeper knowledge of how bilingual education affects different areas of language acquisition. Both constructs (vocabulary knowledge and phonemic discrimination ability) were measured with the accuracy scores and d-prime scores, as a secondary aim of this thesis was to assess what measurement is better suited to explaining the findings. So far, only few studies have explicitly compared these two measures.

2. Theoretical Background

2.1 Content Language Integrated Learning

Content Language Integrated Learning (CLIL) in the Netherlands has seen a steep rise in popularity in recent years. In between 2013-2017, the number of children who were in a bilingual program (in Dutch: TTO *tweetalig onderwijs*, "bilingual education") rose from 29 thousand to 36 thousand (Van 't Erve, 2021). In 2005, only 5 thousand pupils were following bilingual education (Admiraal et al., 2006). In the current thesis, CLIL is used when referring to bilingual education in general, and TTO to refer to the implementation of bilingual education in the Netherlands specifically.

TTO is advertised as a way to ameliorate Dutch children's foreign language skills (usually English) without any disadvantages for their Dutch abilities. Because CLIL relies on additional language exposure through the teaching of content in the foreign target language, and not necessarily on more instruction in language classes compared to non-CLIL pupils, *implicit* language learning is stimulated (i.e., learning from input without being aware of the learning process (de Graaff, 1997; Hulstijn, 2012)). The opposite is *explicit* instruction, which occurs when a pupil is told something about language with the intent of giving them a strategy by which to learn a feature of language (DeKeyser, 2003), for example when a teacher explains a correct pronunciation or the meaning of a certain idiom. CLIL is based on the idea that pupils do not necessarily need to receive more explicit instruction in order to improve their second language skills. Instead, they learn language implicitly through the additional language exposure in content classes, which indeed seems to happen in practice (Lorenzo et al., 2010). Implicit and explicit forms of instruction are both necessary in acquiring a foreign second language in a classroom. Especially for learning to discriminate phonemes, which is one of the language areas of interest in this current thesis, learners appear to perform better when they are explicitly told when they are perceiving phonemes incorrectly (Thomson, 2018).

Generally the advantageous role of TTO for language learning in the Netherlands is confirmed by academic studies (Admiraal et al., 2006; Huijbregtse, 2001; Rumlich, 2018; Verspoor et al., 2015). Pupils who follow a bilingual education are normally found to have a larger vocabulary size in the target language (Admiraal et al., 2006; Bulon & Meunier, 2020; Jiménez Catalán & Agustín Llach, 2017; Olsson, 2021), better comprehensive reading skills (Huijbregtse, 2001), oral ability (Huijbregtse, 2001; Ruiz de Zarobe & Lasagabaster, 2010) and more skills in various grammatical constructs (Hendrikx, 2019; Ruiz de Zarobe & Lasagabaster, 2010) in comparison to pupils following a monolingual education. Additionally, CLIL pupils were often found to be more motivated to acquire the foreign language (Mearns, 2016; Seikkula-Leino, 2007), especially in secondary education (De Smet et al., 2019). Note that this may also be the reason that they are drawn to CLIL in the first place (Lasagabaster & Doiz, 2017). Furthermore, it has often been pointed out that CLIL pupils are not disadvantaged in how well they learn content (e.g., history, geography) despite the subject being taught in a foreign language: CLIL pupils generally perform similarly to non-CLIL pupils in all non-language subjects (Admiraal et al., 2006; Lorenzo et al., 2010).

However, there are some other aspects of CLIL to consider as well. Bruton (2011) remarks that bilingual education may not necessarily be beneficial in all aspects for pupils, and additionally points out several methodological issues with research that has concluded an advantageous role of CLIL on pupils' language acquisition. Many studies on the CLIL context found that pupils who follow a bilingual track often start out with a higher language proficiency and/or motivation to learn languages in comparison to pupils who follow a monolingual track (Admiraal et al., 2006; Alonso et al., 2008; Pérez Cañado, 2018; Ruiz de Zarobe & Lasagabaster, 2010; Verspoor et al., 2015). It is important for researchers that such initial proficiency differences are not confounded with a beneficial influence of CLIL, especially if the language proficiency of CLIL pupils does not improve more strongly over time in comparison to non-CLIL pupils. A higher proficiency level, in this case, does not indicate an advantageous effect of CLIL on language skills, but rather suggests that CLIL attracts pupils who are already have more language skills. Furthermore, CLIL is not necessarily implemented in the same way across countries, or even *within* the same country (Olsson, 2021), and this may affect the learning outcomes per school. Another matter to keep in mind when it comes to assessing the advantage of bilingual education is that the improvement on language abilities may simply reflect the effect of additional hours of exposure. Several studies report that pupils in a CLIL-track receive more hours of explicit language lessons than non-CLIL pupils, and thus not just additional language exposure through the integration of a foreign language in content lessons. Bruton (2011) notes that these additional hours of language classes seem to lead to a rather small gain in language skills, and that it should be considered whether such a slight improvement is worth the hours of extra effort that pupils put in. Furthermore, it should be noted that quality of exposure has also been found to be a significant factor in language acquisition (Paradis, 2011; Uchihara et al., 2022), and that a lack of improvement may also have to do with a low quality of language exposure instead of the amount of language exposure. In general, if the quality of exposure is sufficient for allowing pupils to improve their language skills, it may be expected that the larger amount of second language (L2) language exposure in CLIL yields a higher proficiency level for pupils. Furthermore, CLIL pupils' level of language improvement can be expected to occur at a faster rate than that of a non-CLIL pupil because of this additional language exposure. Verspoor et al. (2015), for instance, report that throughout the first year of high school, CLIL pupils do improve their writing and vocabulary skills significantly better compared to non-CLIL pupils, adding that the CLIL pupils receive seven times the amount of exposure and have an additional three hours of explicit English instruction per week (a total of five hours of instruction versus only two hours for non-CLIL pupils).

Explicit instruction, which does not typically occur outside of language classrooms, is often considered to be required in order to acquire or improve upon certain forms of linguistic knowledge (Lacabex & Gallardo-del-Puerto, 2020). This type of instruction is something that may not occur ordinarily within CLIL contexts, as it has been remarked that some teachers pay little attention to language in content lessons (Olsson, 2021) and teachers may rely on implicit acquisition for pupils (Rumlich, 2018). This may be because some of the content teachers' limited language proficiency (Bruton, 2011) and their inability to stimulate language education to its full potential (Huijbregtse, 2001). However, the proficiency level of teachers is highly variable and no general remarks can be made about their level. Some studies, in fact, find that CLIL teachers do seem to be aware of the status of language and report to using adapted pedagogies for bilingual education (Van Kampen et al., 2018). This only highlights how different the implementation of CLIL can be across schools.

Some studies do find that CLIL pupils do improve to a significantly stronger degree in comparison to non-CLIL pupils in language areas such as vocabulary size (Olsson, 2021; Pérez Cañado, 2018; Verspoor et al., 2015), writing skills (Verspoor et al., 2015) and general linguistic competence including some measures of writing skill, oral competency, and vocabulary; see Lorenzo et al. (2010) and Ruiz de Zarobe and Lasagabaster (2010). There are more studies claiming an advantage for CLIL pupils over non-CLIL pupils in certain areas of language learning, such as receptive word knowledge and reading comprehension (Admiraal et al., 2006), lexical production and writing fluence (Agustín Llach, 2017), and use of phraseological units (Bulon & Meunier, 2020), as they found CLIL pupils to perform better than non-CLIL pupils. However, in these studies, the stronger improvement of CLIL pupils over non-CLIL pupils is either not discussed or is not actually statistically significant, which once again suggests that it is not necessarily bilingual education that causes CLIL pupils to outperform non-CLIL pupils in language skills. Rather, it may be CLIL pupils' motivation and their possible strong language aptitude that causes them to have better language skills. Additionally, these factors may attract pupils to CLIL. A higher language proficiency does not necessarily prove that CLIL has a beneficial influence on pupils' language skills. In order to reach that conclusion, the results of a study should demonstrate that pupils following CLIL improve more strongly than non-CLIL pupils.

Little research has been done so far about the effects of CLIL on phonetic learning, apart from some research on speaking ability. Lacabex & Gallardo-del-Puerto (2020) conducted a study involving vowel reduction in English unstressed syllables and investigated whether or not CLIL pupils were aware of these. Their findings revealed that the CLIL pupils required explicit training in order to acquire this phonological awareness. This is in line with research on phonetic acquisition in the L2 outside of the CLIL classroom (e.g., Kissling, 2015).

To my knowledge, no other studies on the perceptive phonetic abilities of CLIL pupils exist. In order to address this research gap, one of the main aspects of this thesis is to analyse the effect of TTO on pupils' phonemic discrimination ability. Phonetic acquisition is considered one of the hardest parts of language learning (Baker & Trofimovich, 2005) and the lack of research is thus surprising, since it is well-known that Dutch speakers struggle with certain phonemic discriminations in English (Broersma & Cutler, 2008). Research on CLIL has increasingly dealt with vocabulary size, acquisition of grammatical aspects of the foreign language, and reading and writing skills. The lack of focus on pupils' phonetic acquisition seems an oversight that, in my opinion, ought to be corrected in order to gain a better view of the effects of CLIL on language proficiency as a whole. As long as such an important area of language acquisition is under-researched, we do not have a complete idea of how CLIL influences pupils' general language abilities.

2.2 Phonemic discrimination

This section will discuss phonemes and phonemic discrimination, including what they are (Section 2.2.1), what factors affect the acquisition of phonemes in both first and second language acquisition (Sections 2.2.2 and 2.2.3), and the phonemic contrasts that have been chosen for this study (Section 2.2.4).

2.2.1 Phonemes and minimal pairs

Phonemes are the sounds that make up a language and that cause a change in meaning if one phoneme is swapped for the other. However, a change in sound does not necessarily equate a change in phonemes. That is, language is capable of having *phonetic* variation (i.e., the production of one phoneme may differ slightly depending on various factors, and result in different *phonetic variants* of the same phoneme, which are called *allophones*) that does not lead to a *phonemic* change (i.e., a change in the meaning of the word). A phonetic change in one language can be a phonemic one in another. Words between which only one phoneme varies are called *minimal pairs*: examples in English are for example “bat” and “bet”, or “lead” and “mead”. The minimal pair “bat” and “bet” in English is created because of the phonemic difference between the vowels /æ/ (in “bat”) and /ɛ/ (in “bet”).

The ability to distinguish between phonemes is what is called *phonemic discrimination ability*. Discriminating between one's native-language phonemes is easily done; however, in a foreign language it is considerably harder. Such difficulties cause issues with both speech production and comprehension. To return to the "bat"/"bet" example mentioned above: in Dutch, the /æ/ vowel does not exist. Any use of this speech sound in Dutch would be perceived as a *phonetic* variant of the /ɛ/ vowel, which is often how native speakers of Dutch categorise the /æ/ sound in English (Weber & Cutler, 2004). When learning English, the acquisition of this new phoneme that is not in the native-language inventory is difficult for native speakers of Dutch (Broersma, 2005; Díaz et al., 2012; Thorin et al., 2018; Weber & Cutler, 2004). Unfamiliar phonemes are often placed in a phonemic category that a speaker knows from their native language before they acquire the contrast, and thus misheard as a phoneme with which they are familiar. Therefore, the /æ/-/ɛ/ minimal pair was chosen as one case of investigation for the present study.

2.2.2 Age effects in the acquisition of phonology

Phonemes in a speaker's native language are acquired early in life. Infants start off being able to discriminate differences in speech sounds that adults no longer can, and have honed in on the speech sound in their mother tongue when they are only one year old (Kuhl, 2004). This allows them to learn their first language(s) (L1) well, but also means they are less able to discriminate speech sounds in non-native languages from a young age (Kuhl, 2004). It is well-known that comprehending speech in one's second language – especially when this language has been learned after puberty – can be rather difficult, even for more fluent speakers. That is, speech can be fast, unfamiliar speech sounds can be encountered, and unfamiliar accents may shift the pronunciation of some previously-understood words until they are nearly unrecognisable to the L2 speaker's ear (Weber & Cutler, 2004).

The critical period (i.e., the time in one's life one is most capable of acquiring a certain language phenomenon) for acquisition of a native-like phonemic inventory is often considered to be somewhere between the ages of 3 and 12 (Abu-Rabia & Kehat, 2004). The existence of a critical age in its strict sense implies that after this age has been reached, phonological acquisition is considered to be difficult. Furthermore, it implies that age plays no role beyond this threshold (i.e., the degree of difficulty should be the same for learners in their twenties as for learners in their fifties). Some studies, however, found that phonology will gradually become harder to acquire with age even after puberty has ended (Saito, 2015; Schepens et al., 2022), suggesting that age-related difficulties do not halt after a possible critical age. Whichever theory may be the right one, this early phonemic acquisition may cause difficulties for learning the phonology of a further language later in life. Moreover, these difficulties that are often encountered in adult phonological acquisition explain why phonemic acquisition is considered to be strongly affected by age effects (Schepens et al., 2022). Even many young learners (i.e., under age 12) do not exhibit a nativelike phonological behaviour, and although some older learners do learn the phonology of a foreign language to a high level, the probability of them acquiring this area of language to that level is low (Abrahamsson & Hyltenstam, 2009). In sum, we are not entirely sure when age stops affecting phonological acquisition, or if it stops at all. However, we do know that age is an important factor in phonological acquisition, with younger learners often acquiring a more native-like phonological proficiency in comparison to older learners (i.e. during or after puberty).

2.2.3 Other factors that influence the L2 acquisition of phonology

Several factors other than age have been found to influence second language phonological acquisition, but it remains an intricate story, as many foreign language learners show a high degree of individual variation in their phonetic abilities (Darcy et al., 2015; Kissling, 2014). One factor that is often found to be influential is the learner's L1: a speaker's perception and production of speech sounds in a foreign language will be influenced strongly by their native language in the acquisition of a foreign language (Darcy et al., 2015; Van Leussen & Escudero, 2015). Non-native accents are prevalent even in highly proficient speakers (Wolfswinkler & Reinisch, 2016), as a native language can impede a learner's ability to form new phonemic categories (Baker & Trofimovich, 2005; Bosch et al., 2000). That is, we are so attuned to the phonemes in our native language that our existing phonemic categories can make it difficult to perceive and form new ones.

Besides L1 and age, quantity and quality of language exposure have been found to be important for acquiring phonetic information (Paradis, 2011; Saito, 2015). That is, more language exposure, as long as this exposure is qualitatively adequate for pupils to acquire new phonemic contrasts, leads to better learning outcomes. More exposure to a language allows learners to gain more experience in both perceiving and producing phonemes, which naturally causes them to improve more strongly over learners with less exposure (and therefore less practice).

Additionally, a factor that impacts phonological acquisition is language aptitude. Individuals who have been found to reach high levels of phonological proficiency as late learners were more likely to have a high language aptitude. Thus, these learners have the ability to acquire a foreign language phonology because they are adept at processing phonetic detail and are better able to discriminate phonetic differences between their L1 and L2 (Kissling, 2014). Language aptitude cannot be taught, but is a factor that varies per individual. It has been found that language aptitude is not a significant factor in differences in language proficiency outcomes for young learners, but it is important for learners who start acquiring a target language during puberty. The late learners with a high aptitude were eventually more successful in language acquisition in comparison to late learners with a lower language aptitude (Harley & Hart, 1997).

Another factor that is often found to correlate with a better phonemic discrimination ability is vocabulary size. Georgiou et al. (2020), for instance, found that L1 Russian speakers with a larger vocabulary in their L2 English were more attuned to the acoustic cues in the experiment. As a result, these speakers were better at discriminating vowel contrasts in English than speakers with a smaller vocabulary size. Another study by Daidone and Darcy (2021) concurs with these findings, having tested the perception of four phonemic contrasts in L2 Spanish that do not occur in the native English language of the participants. In their study, vocabulary size in the L2 was considered to be the most accurate predictor of the ability of English speakers to remember the correct phoneme in a word. They found that vocabulary size was a significant factor in three of the four tested phonemic contrasts. Phonological short-term memory was a significant predictor for the one remaining phonemic contrast.

Furthermore, Daidone and Darcy (2021) theorise that learning a higher number of words that are phonemically similar requires the phonetic representation to be more detailed in order to differentiate between these words. When lexical representations are stored in the mind with phonetically incorrect information, this leads to an incorrect pronunciation in conversation, even though speakers may technically know the acoustic differences between the phonemes and even be able to produce them (Hayes-Harb & Masuda, 2008; Llompart, 2019; Llompart & Reinisch, 2019). This would explain why some speakers are able to imitate words accurately or produce a native-like pronunciation well when reading, but fail to do so in conversations. Llompart and Reinisch (2018) add to this that memorisation of less-accurate phonetic detail in the L2 can not only entail pronunciation issues, but can also lead to more difficulty in word recognition. However, not all research unanimously agrees with the correlation between vocabulary size and phonemic discrimination. Llompart (2021b) found that a larger vocabulary size correlates with the ability to discriminate /æ/-/ɛ/, but *only* in highly proficient speakers. Darcy et al. (2015) did not find a correlation between vocabulary size and phonological acquisition at all for L1 Korean speakers learning L2 English. Instead, they found that the ability of L2 English speakers to acquire native-like phonemic discrimination relied on working memory capacity most strongly. As earlier studies have not found the same results pertaining to the correlation between vocabulary size and phonemic discrimination, the present study's inclusion of vocabulary size as an independent variable might provide us with more insight on how these two language skills are correlated.

2.2.4 Phonemic contrasts in the current study

In the current study, pupils' ability to discriminate several phonemes is measured with item pairs consisting of one word pronounced correctly and one manipulated version of that word in which the phoneme has been swapped with its counterpart. Five phoneme contrasts are tested: /b/-/p/, /v/-/f/, /d/-/t/, /z/-/s/ and lastly /æ/-/ɛ/. It is expected that /æ/-/ɛ/ is the phoneme discrimination that pupils will have most difficulty with considering the absence of this phonemic contrast in Dutch (see Section 2.2.1). German, similarly to Dutch, lacks the /æ/-/ɛ/ phonemic distinction: both languages tend to assimilate any instance of /æ/ to the closest phonemic category in the native language, the mid-front vowel /ɛ/. Llompart & Reinisch (2017) found an asymmetry in the /æ/-/ɛ/ perception in German speakers: words containing /ɛ/ were more easily recognised than words containing /æ/. An additional difficulty in the /æ/-/ɛ/ discrimination is the opacity of the English orthography; the inconsistent spelling of these phonemes may hinger perception for L2 learners. Yet, words with the /æ/-/ɛ/ contrast (such as "bat" and "bet") are usually not stored as homophones in the L2 mental lexicon. Rather, L2 speakers are aware of the different phonemes but have a harder time perceiving [æ] as it activates both words including the /æ/ and /ɛ/ phonemes, whereas the production of [ɛ] only activates /ɛ/ (Llompart, 2021a; Weber & Cutler, 2004).

Another reason why the /æ/-/ɛ/ distinction may be hard to perceive is that it has a rather low phonetical salience; that is, the contrast may be hard to notice in the input. (Lersveen, 2018). However, Hommel (2018) does not see saliency as the most important factor to determine a contrast's difficulty, and argues that frequency of a phonemic contrast, rather than its salience, has a more important role in determining the learning outcome. In her study, she found that the /æ/-/ɛ/ distinction was not difficult to acquire for Dutch speakers at all. She argues that the frequent occurrence of this particular phonemic difference allows Dutch speakers to notice this contrast, and their awareness subsequently benefits their ability to phonemically discriminate between these sounds. However, it may also be the case that exposure may not be enough for L2 learners to learn to distinguish between the vowels but that explicit instruction on the /æ/-/ɛ/ phonemes is necessary to acquire this discriminatory ability (Lacabex & Gallardo-del-Puerto, 2020; Llompart & Reinisch, 2017).

The other English phoneme pairs that are to be discriminated in this study are the voiced and unvoiced labiodental fricatives (respectively /v/ and /f/), the voiced and unvoiced alveolar fricatives (/z/ and /s/), the voiced and unvoiced bilabial plosives (/b/ and /p/) and lastly the voiced and unvoiced alveolar plosives (/d/ and /t/). All the consonants will be tested in word-final position. These consonants are all present as phonemes in Dutch. Their production in English and Dutch may slightly differ, but the Dutch language does have the voiced-unvoiced distinction for these phonemes, which may make it easier for the participants to discriminate between them. That is, even if the sound is not produced *exactly* the same in both languages, it is similar enough for Dutch speakers to have mapped that speech sound to their native phoneme category in their mental lexicon in the L2. Because of this reason, one can expect the consonant-contrasts in this study to be less problematic for learners than the vowel contrast. One difficulty that the participants may encounter is because of word-final devoicing in Dutch. Dutch devoices consonants at the end of a word, but English does not. Possibly, the participants may find it difficult to perceive voiced consonants at the end of a word.

Furthermore, it is expected that pupils will perform best at discriminating between items with plosives and only second-best at discriminating fricatives, as Dutch is currently undergoing a devoicing of the fricatives (De Schryver et al., 2013; Pinget et al., 2020; Van de Velde & Van Hout, 2001). In the Netherlands, both /v/ and /z/ have been undergoing a strong devoicing process for years, causing a production of unvoiced fricatives in the place of voiced ones (e.g., *huisen** (huizen), *lefer** (lever)) (De Schryver et al., 2013; Pinget et al., 2020; Van de Velde & Van Hout, 2001). This process of fricative devoicing may cause pupils to perform less accurately on discrimination between voiced and unvoiced fricatives than on the discrimination between voiced and unvoiced plosives.

2.3 Lexical decision ability

Lexical decision tasks are often used to measure receptive vocabulary knowledge. In a lexical decision task, participants are presented with either existing words or non-words and asked for each item to either reject or accept it as a real word. It is expected that a larger vocabulary size leads to better performance on such tasks, as knowing more words will likely lead to a better ability to correctly accept words and reject and non-words. It is generally found that CLIL pupils have a larger vocabulary size (Admiraal et al., 2006; Agustín Llach, 2017; Bulon & Meunier, 2020). In this thesis, the lexical decision tasks measures both phonemic discrimination ability as well as lexical decision ability. In this way, both abilities are evaluated by using the same task to see how well pupils do in these language areas.

Broersma (2012) found that native and non-native listeners process lexical information differently: L2 speakers respond more slowly and less accurately on lexical decision due to the lexical competitors that are activated when their phoneme categories are less robust. Furthermore, L2 speakers' comprehension also involves the activation of words in the L1, contributing to the larger number of lexical competitors. Broersma (2012) found that Dutch speakers, for instance, have a tendency to mentally activate minimal pairs with /æ/ upon hearing [ɛ] and that they are more likely to accept an incorrect word. For example, if Dutch speakers hear [kɛt], they are likely to activate "cat" despite the correct pronunciation being [kæt]. In accepting [kɛt] as a possible pronunciation for "cat", they are incorrectly accepting a non-word (or, as these words with the target phoneme altered are referred to in this study, a near-word). In this way, lexical decision tasks can be used to not only measure receptive vocabulary knowledge but also provide us with information about participants' ability to discriminate phonemes. Llompart (2019) similarly uses the lexical decision task to collect data to understand the connection between phonetic ability and the memorisation of words in participants.

Lexical decision tasks may thus have several uses: they allow for measuring receptive vocabulary size and for measuring whether participants can recognise phonetic detail. Note that the lexical decision task is not always used for measuring phonemic discrimination, as some lexical decision tasks do not use auditory input (i.e., some lexical decision tasks use written input), but for this study this task is a useful tool for measuring phonemic discrimination ability. Because vocabulary size and phonemic discrimination are measured with the same task, it allows us to compare the performance on both components of the task. If the vocabulary size of the pupils is too small, they will do poorly on both components. However, it is expected that pupils will have more difficulty with the phonemic discrimination part of the task compared to the lexical decision part. By using the same task to measure both vocabulary size and phonemic discrimination ability, we can see whether errors on the phonemic discrimination component are due to difficulty with their vocabulary knowledge. Alternatively, pupils may perform well on the lexical decision part of the task but perform relatively worse on the phonemic discrimination part, which would indicate that it is only phonemic discrimination ability with which they struggle.

3. The current study

This thesis aims to address the research gap regarding the effect of bilingual education on phonemic discrimination ability in the L2, and to further extend research on the effects of bilingual education on vocabulary knowledge. In order to do so, this thesis focuses on the performance on an auditory lexical decision task, which measures receptive vocabulary knowledge as well as phonemic discrimination ability. The test had one section to measure vocabulary size with items that were either real words or non-words. The other section measured phonemic discrimination by including item pairs: the same word occurred twice in the test, one time pronounced correctly and one time with a target phoneme (either one of the consonants or the vowel of interest) shifted to be pronounced incorrectly.

The participants were Dutch-speaking pupils (12 to 16 years old) of the Dutch high school the Varendonck College in Asten. The pupils followed either a TTO program (involving immersion in L2 English) or had chosen to follow a monolingual (i.e., fully Dutch) program. This study intends to address whether TTO pupils had an advantage over non-TTO pupils with respect to phonemic discrimination ability and vocabulary knowledge in the L2 English, and whether this advantage increased over the years or, alternatively, whether TTO and non-TTO pupils acquired their L2 language skills at the same pace. We gain insight in receptive vocabulary size as pupils reject or accept items that are either real or non-words. In this thesis, this skill is referred to as "lexical decision ability", which is not a term that is ordinarily used in research. It refers to vocabulary size, but two other additional vocabulary tests were administered which are used as covariates in some of the analyses that have been conducted. For clarity's sake, the scores on the lexical decision task that measures vocabulary size will therefore be referred to as *lexical decision ability*. The scores of the two additional vocabulary tests (the Dutch and English PPVTs) are referred to as vocabulary size.

We also measured phonemic discrimination ability with the lexical decision task, as pupils either rejected or accepted items that were real words or near-words (i.e., real words in which one target phoneme has been altered, resulting in a word that does not actually exist). The TTO and non-TTO pupils all took the same three tests (the lexical decision tasks and the Dutch and English PPVTs). It was analysed whether differences in TTO-track and school year (Year 1, 2 or 4) were significant factors in pupils' scores lexical decision ability (as a proxy of vocabulary knowledge) and phonemic discrimination ability of vowels, fricatives and plosives. The following research questions were formulated:

Main research question

What is the effect of bilingual secondary education on Dutch high school pupils' lexical decision ability and phonemic discrimination ability in English?

Sub-questions

RQ1. Is there an effect of school year and language track on performance on the lexical decision part of the task? Is there an interaction effect between school year and language track?

RQ2. To what extent are lexical decision ability and performances on the other two vocabulary tests (PPVT English and PPVT Dutch) related?

RQ3. Is there an effect of school year and language track on performance on the phonemic discrimination component of task? Is there an interaction effect between school year and language track?

RQ4. To what extent are phonemic discrimination ability and performance on the two vocabulary tests (PPVT English and PPVT Dutch) related?

RQ5. To what extent does phonemic discrimination ability depend on manner of articulation (vowels, fricatives, plosives)?

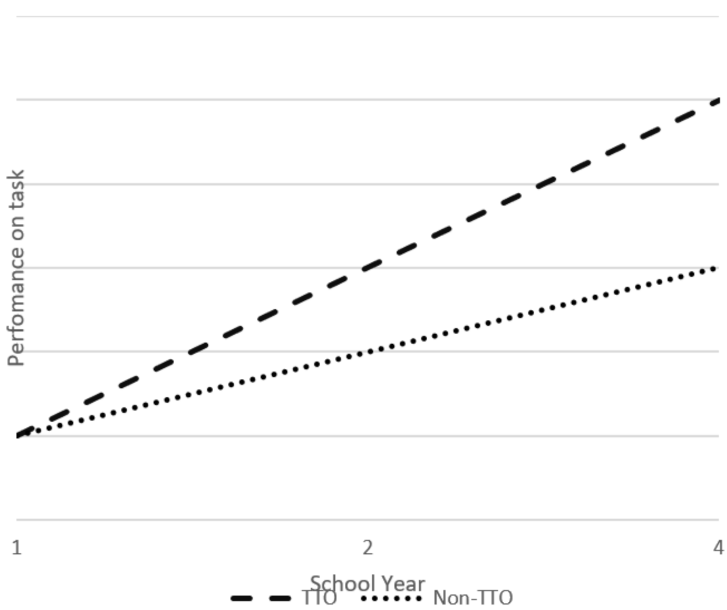
RQ6. Do the results of the analyses reveal that d-prime leads to a better fit of the statistical models than accuracy scores?

The vocabulary tasks (the PPVTs) measure a similar construct as the lexical decision component of the task in this design, except for the fact that the vocabulary task is administered in both Dutch and English whereas the lexical decision task was only in English. The vocabulary task is thus a separate assessment and will be used as an independent variable in the analyses, while the performance on the lexical decision component (*lexical decision ability*) and the phonemic discrimination component of the lexical decision task will be used as the dependent variables.

Hypothesis regarding the research questions

If TTO does indeed affect second language learning advantageously in comparison to traditional language teaching (as a possible answer to RQ1 and RQ3), we would expect an interaction effect between school year and language track: the lexical discrimination and phonemic discrimination abilities of the TTO pupils may be expected to increase at a significantly faster pace from their first year of high school to the fourth year compared to non-TTO pupils. This could lead to a wider gap in both linguistic abilities between fourth-year TTO and non-TTO pupils compared to first-year TTO and non-TTO pupils. Figure 1 provides an example of what this interaction may look like.

Figure 1
Expected interaction between school year and TTO



The performance on the PPVT vocabulary tests (RQ2 and RQ4) are expected to be related to the performances on both parts of the lexical decision task, especially the English PPVT considering the lexical decision task is also in English. Since lexical decision ability also measures vocabulary size, just with a different task, a high correlation is expected. The correlation with phonemic discrimination ability is predicted to be significant, as it is expected that a pupil being more skilled in one language area (in this case, vocabulary knowledge) makes it more likely for them to have a possible advantage in another language area (i.e., phonemic discrimination ability) as well.

The analysis comparing the phonemic discrimination ability on vowels and consonants may also offer interesting insight (RQ5). The consonant contrasts used in this thesis may be more easily perceived than the vowel contrast since these are all native phonemic contrasts in Dutch (see Section 2.2.4). However, consonants are devoiced in Dutch word endings whereas this is not a phonotactic constraint in English, and this may lead to difficulty in rejecting a near-word in which the voiced and unvoiced items have been exchanged for Dutch pupils. As for the vowel contrast, the /æ/-/ɛ/ distinction is not phonemic in Dutch, and pupils may find the distinction difficult to perceive. If it is perceived at all, the realisation of /æ/ as [ɛ] may be accepted, as /ɛ/ is the closest categorised phoneme in the Dutch sound system. Possibly, the additional exposure that TTO pupils have had will allow them to have acquired the /æ/-/ɛ/ distinction, which non-TTO pupils have not received and thus may struggle with. However, all pupils may be able to hear the difference between voiced and unvoiced consonants since this is a phonemic contrast that also exists in their native language of Dutch. This is why the hypothesis is that TTO pupils will outperform non-TTO pupils on phonemically discriminating the vowel contrasts but that they will score similarly on the consonant contrasts. It is also expected that higher school years will score significantly better than lower school years, because of the additional exposure that pupils in higher school years have received. Another possibility may be that teachers in the TTO-track may not have offered enough explicit instruction for the TTO pupils to acquire these phonemic contrasts and they might thus not perform any better than the non-TTO pupils on any of the phonemic contrasts. Additionally, teachers may not offer language exposure of sufficient quality to help pupils improve their ability to discriminate phonemes. Little research has been conducted about the advantage of bilingual education in the improvement of phonetic skills, so it is hard to know whether TTO offers enough exposure for pupils to gain an advantage in phonemic discrimination ability. This thesis strives to address this research gap.

The final research question (RQ6) aims to compare the use of the d-prime scores versus accuracy scores as outcome variables of the lexical decision task. The d-prime is a score that is calculated from the “hits” (correct answers on real-word items) and “false alarms” (incorrect answers on non-word items (lexical decision) or near-word items (phonemic discrimination)). It is meant to measure a participant’s sensitivity to the signal: i.e., the score on the d-prime indicates one’s awareness to discriminate between correct and incorrect items (Haatveit et al., 2010). The d-prime takes away the response bias (e.g., a participant saying all incorrect words are correct because they cannot discriminate phonemes) in a way that accuracy scores do not. Therefore, we hypothesise that the d-prime is the preferred outcome variable in an analysis, as it should prove more correct to a participant’s actual skill. Many studies still use accuracy scores instead of d-prime scores. The d-prime and accuracy scores and their issues are explained in more depth in Section 4.3.2.This thesis measures both in order to compare what the differences are and what outcome variable (i.e., d-prime or accuracy) fits the statistical models better.

4. Methodology

4.1 Participants

The data of the present study were collected over the course of three academic years (2014-2015, 2015-2016 and 2016-2017), with data collection sessions taking place once per year, at the Varendonck College in Asten, the Netherlands. The Varendonck College is a Dutch high school that offers a monolingual track (only Dutch) as well as a bilingual track (Dutch and English: also called TTO *tweetalig onderwijs*, "bilingual education"). The Varendonck College has offered the TTO-track since 2009-2010. This TTO-track is only offered to VWO-level pupils. In the Netherlands, VWO stands for *voorbereidend wetenschappelijk onderwijs* which translates to "preparatory scientific education" and is generally intended to prepare its pupils to follow a higher education at a Dutch university. The VWO-level education at a Dutch high school takes six years to complete. Pupils start in Year 1 when they are 12 years old and they will be 18 years old by the time they finish. The vwo-pupils of the Varendonck college are free to choose between the TTO-level and the monolingual Dutch track. Therefore, all pupils tested in the present study were VWO-pupils.

In the first three years of the TTO-track offered at the Varendonck College, about 60% of all education is in English, spread over several courses (language classes and content courses). In order to prepare the pupils for the final high school exams which are in Dutch, the amount of English considerably decreases in the final three final years of their education, to only 30%. Furthermore, the pupils need to have completed a two-week long internship in an English-speaking environment by the end of their sixth year.

A total of 172 pupils, all native speakers of Dutch, were tested for this study across three separate school years. These were Year 1 (12-13 years old), Year 2 (13-14 years old) and Year 4 (15-16 years old). Of these pupils, 133 were in the TTO-track and 39 were in the monolingual (non-TTO) track (see Table 1). One pupil in Year 4 of the TTO-track was not included in final analyses because a response bias was detected in their answers on the lexical decision ability part of the task (see Section 4.2.2 for an explanation on how lexical decision was measured; see Section 4.3.1 for an explanation on how the response bias was detected). It should be noted that the number of non-TTO pupils is considerably lower than the number of TTO pupils: this may lead to difficulties finding effects in the statistical analyses. Furthermore, 17 teachers were tested. It was decided that the teachers would not be included in any analysis for this thesis, however, as the focus lies on the pupils' (and not the teachers') phonemic discrimination ability and lexical decision ability. An analysis of the role of teachers' input on pupils' language skills would be interesting for a future study.

Table 1
Pupils and their distribution over school years and TTO/non-TTO track

School year	TTO	Non-TTO	Total
Year 1	52	14	66
Year 2	40	10	50
Year 4	40	15	56
Total	132	39	171

4.2 Lexical decision task

The lexical decision task was used to measure two things: (1) phonemic discrimination ability and (2) lexical decision ability (as indicatory of vocabulary knowledge), both in English. The task involves the participant hearing a word and choosing whether this word exists or not. The scores on the two parts of this task are the dependent variables of the present study. The same task was used by Broersma and Cutler (2008). Several item types were tested in their task: non-words, real words, and *phantom words*, which are real words in which a phoneme has been manipulated (in this thesis called near-words; e.g., *groof** from *groove*). Their experiment found that *phantom words* can activate the original real word in English for learners of L2 English during listening (e.g., hearing *groof** can activate the word *groove*). In the current thesis, the task is not used to investigate whether phantom words activate their real counterparts but rather as combination of a phoneme discrimination task measuring phonemic discrimination ability and a lexical decision task measuring lexical decision ability. The task consists of 200 words in total that are tested (the same non-words, real words and near-words that Broersma and Cutler (2008) used). The procedure for testing is explained below in Section 4.2.3.

4.2.1 Phonemic discrimination

The phonemic discrimination part of the task included 64 *target* items, which were the items that in the original experiment by Broersma and Cutler (2008) were meant to assess participants' activation of words in their L2 English and which the present study used to test phonemic discrimination ability. Appendix 1 contains the word list for the target items. Half of these items were monosyllabic English words which had a word-final consonant /p/, /b/, /d/, /t/, /f/, /v/, /z/, or /s/. The other half was monosyllabic English words with either the stem vowel /æ/ or /ɛ/. Half of the target items were phonetically manipulated and were devised so that the final-word consonants became voiced when they were devoiced and the other way around (e.g. /d/ to /t/: *glide* to *glite**) or the vowels were exchanged (/æ/ to /ɛ/ or vice versa: for example *rank* to *renk**). Importantly, this manipulation thus involved a phoneme switch and resulted in an incorrect pronunciation of the words (and are referred to as *near-words* in this current study). The other half of the target items consisted of the original, correct counterparts of the manipulated words (the real words). The near-word and real word together form an *item pair*, and were analysed together (see Section 4.3.2). Pupils were exposed to only *one* item of the item pairs during the task, which means there were two versions of the lexical decision task.

These item pairs were used to measure phonemic discrimination ability, as it was assessed whether the participants were able to discriminate between the (incorrect) near-words and the (correct) real words. The aim was to investigate whether participants would correctly accept a real word, which was pronounced correctly, and whether they would correctly reject a manipulated near-word. If they reject real words and/or accept near-words, this is taken as an indication that the pupils are unable to discriminate the phonemes correctly. Of course, rejecting correct words may also be a sign that they do not know the word. This is why the *filler* items are also analysed, which are used to measure lexical decision ability in the present study.

4.2.2 Lexical decision

The remaining 136 items are two categories of filler items. In Broersma and Cutler's (2008) study these were not analysed and were used in between target items, but in the present study they are used as an indicator of lexical decision ability (indicating vocabulary size). Of the filler items, 68 items were real existing words in English, and the final remaining 68 words are non-words in English (e.g., *fraf**, *noik**). The word list for the filler items can be found in Appendix 2. The filler items were used to test lexical decision ability, since these were real words and non-words, and thus will inform us about pupils' vocabulary size. In this case, the existing words should be acknowledged as correct and the non-words should be acknowledged as incorrect: if pupils perform badly on this part of the lexical decision task, it suggests that they have a small vocabulary since they do not recognise existing words and/or are unable to reject non-words. The filler words are also used as a way to control for the phonemic discrimination part of the task: if the scores on the filler items are lower than on the target items (measuring phonemic discrimination), this would indicate that there is an issue with the pupils' vocabulary size and that they might reject correct words on the phoneme discrimination part of the task simply because they are unfamiliar with these words. However, the expectation in the present study is that the filler items will reveal that pupils are capable of recognising real words and identifying non-words and that the scores on phoneme discrimination will be significantly lower.

4.2.3 Procedure for the lexical decision task

Before the lexical decision task, the pupils received task instructions in Dutch. After the instructions, the pupils went into a quiet room in groups of maximally six participants to start the task. The task was conducted on a laptop and the pupils would wear headphones to hear the items. The instructions would appear one more time on the screen right before the start of the task. The pupils were first familiarised with the task by answering ten practice items, after which a screen appeared to tell the participants to ask any questions at that point in time. If there were no further questions, or after the questions were answered, the actual task with the target and filler items would start. Two versions of the lexical decision task exist, as each participant only heard one item of the each item pairs (i.e., the real word or the manipulated near-word version). The order of items depended on the which of the two versions the participants would do. *All* the filler items (measuring lexical decision ability) were included in the lexical decision task, independent of what version the pupil received for the target items. The items were spoken one at a time in a fixed order for each version.

For each trial, the pupils had a maximum of ten seconds to respond by pressing either 'j' (yes) or 'n' (no), indicating whether an item existed or not. The text 'j' (*ja*, "yes") and 'n' (*nee*, "no") appeared on the laptop screen after the item was played to remind pupils to answer or to show them that their answer had not been registered (i.e., if they pressed a key too quickly after hearing an item). If the pupil did not answer within ten seconds, their answers would be logged as 'timed out' by the system. After ten seconds had passed or after the pupil responded, there was a pause of 800ms before the next item was spoken. After 100 items were completed, the pupil were instructed to raise their hand, so that the researcher could start the second part of the test. The second part started against with one practice item, after which the pupil would continue with the final 100 items. When the pupil was done with all items, they waited in silence for the entire group to be finished with their task.

4.2.4 PPVT

Vocabulary was measured using the Peabody Picture Vocabulary Test (PPVT), which is a vocabulary test originally developed by Dunn and Dunn in 1959. It has been revised several times since (Campbell & Dommestrup, 2010). During the vocabulary test, participants heard stimulus words while being shown four black-and-white drawings, one of which matches the word. The pupils' task was to select the drawing that best corresponded to the word's meaning. In the present study, the PPVT was conducted in both Dutch and English to test pupils' receptive vocabulary size in both languages. The scores can range from 20 to 160.

The PPVT was administered on paper by a researcher or a teacher-researcher who was a native or near-native speaker of the language of the PPVT (either Dutch or English). In any cases this was not possible, the PPVT was administered with a recorded file on a laptop. The pupils joined the administrator in a quiet room in the school to avoid disturbances, during school hours and with a teacher present to supervise. The most recent version of the PPVT, the PPVT-IV Form A, was used for testing English vocabulary. The participants started at the beginning (the PPVT is an adaptive test; the starting place can change depending on the participants' level, but in the current study all pupils started at the beginning) and they would be halted if they made eight mistakes in a set of trials. There were twelve trials in a total of seventeen sets. The Dutch PPVT was begun at the advised starting point for the pupils' age group. The participant was stopped at the same point as in the English PPVT, after making eight mistakes in one set of trials.

4.3 Analysis

4.3.1 Calculating d-prime and accuracy

Accuracy was calculated as the percentage of correct responses. This was transposed to a simple scale of 0 to 1 in which 0 means no correct answers at all and 1 means all answers were correct. However, the risk of a research design such as in this study is that participants have a 50% chance of guessing the right answer when uncertain.

Participants may be biased to form a strategy where they either always answer "yes" or always answer "no" when asked if a word is real for any numbers of reasons: this strategy is called response bias. In order to avoid response bias, the d-prime for all participants and items was calculated. This is done by first calculating the percentage of "hits" for all existing items (i.e., correctly identifying real words as real). Then the same is done for the non-words and near-words and the percentage of "false alarms" of each participant is calculated (i.e., the percentage a participant incorrectly identified non-words or near-words as real). Once these percentages are calculated, they are transformed into z-scores. The d-prime is then calculated by subtracting the z-score of the "false alarms" from the z-score of the "hits" (Haatveit et al., 2010).

As mentioned in Section 4.1, calculating the response bias (as expressed by the d-prime) can lead to participants having to be excluded from analysis. A d-prime of 3 or above is a near-perfect score, whereas a d-prime of 0 or lower suggests that pupils were guessing, likely because the task was considered to be difficult. That is, a d-prime score lower than 0 indicates that participants more often incorrectly answer "yes" than they do correctly (i.e., they claim items that are either manipulated or non-words to be real words more often than they claim that real filler or target words are existing words). This means that pupils whose d-prime is under 0 do poorly at identifying the correct items and/or rejecting incorrect items.

For the target items measuring phonemic discrimination, which are more difficult than the filler items measuring lexical decision ability, this is no reason to exclude the participants. That is, it is expected that pupils who are less able to discriminate phonemes will score negatively on the d-prime. However, the filler words measure vocabulary size rather than phonemic discrimination: the items on this task should be easier to perform well on, as pupils are more likely to be able to recognise words in English than they are likely to be able to discriminate between difficult phoneme contrasts. Indeed, as expected, there was only one pupil whose d-prime falls below 0 on the filler items.

As accuracy does not allow us to detect certain biases, d-prime should be the preferred dependent variable rather than accuracy in statistical analyses. However, some researchers still choose to analyse accuracy scores over d-prime (e.g., Broersma & Cutler, 2008; Lemhöfer & Broersma, 2012), despite the fact that d-prime is able to detect bias. The only advantage of measuring accuracy is that it is generally considered to be easier to understand accuracy percentages, since this is a measurement that is more familiar to most people. Despite the advantages of d-prime, it is not often used in linguistics outside of studies on word recognition tasks (Huibregtse et al., 2002) and it often depends on the researcher and the journal whether d-prime is used or not. In this thesis, both the d-prime and accuracy scores are analysed to see which dependent variable leads to a more robust statistical model. It is expected that the d-prime, as it is sensitive to bias and presumably is a more accurate report of pupils' abilities, will lead to a model that is better able to explain the data.

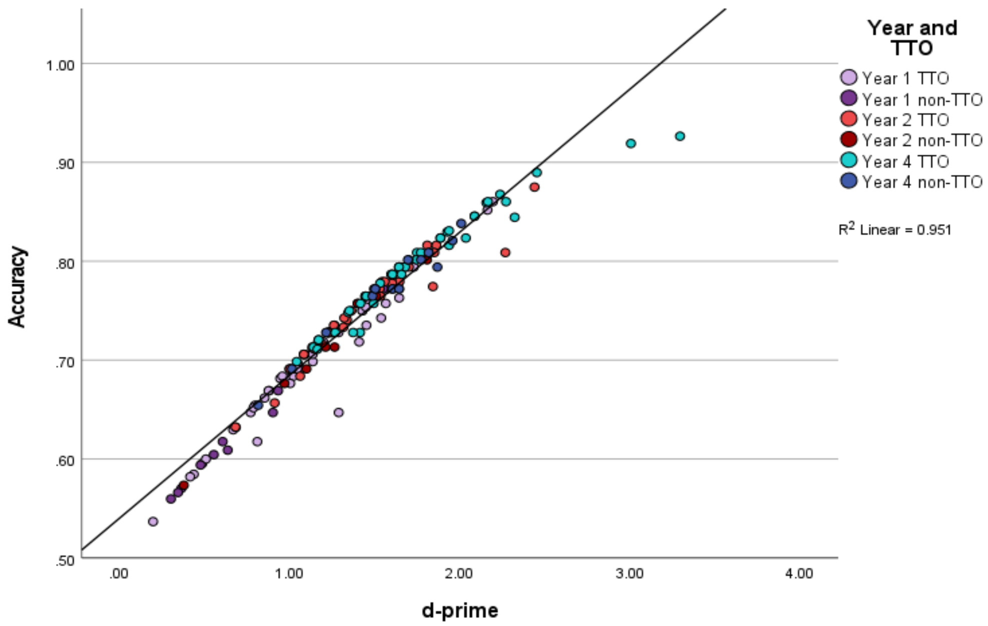
4.3.2 Data cleaning

The dataset used for the present study was collected several years ago (2014-2017). All raw data of the task was handed down by researchers originally involved in the project and required to be explored to assess how the data had been organised and what analyses was required. The final PPVT scores (of both the English and Dutch PPVTs) were all in a single file for all pupils, so little needed to be examined for the analysis of these scores, as they only included one score per participant. The answers on all the task items were collected in one file (i.e., for both the target items measuring phonemic discrimination ability as well as the filler items measuring lexical decision ability). In this file, errors were found in the coding of several answers (a correct answer coded as incorrect or the other way around). These errors were all found with a filter in SPSS and removed from the dataset, since there is no reliable way of knowing where the error originated. The practise items and all items that were not answered in time were also deleted from the analysis. Across all pupils, a total of 155 incorrectly coded trials were not considered for the analysis, as well as the 1910 practise items and 38 timed-out trials. For the vocabulary part of the lexical decision task, all single *items* that were either incorrect or correct were analysed. This means that *all* the items that the pupils answered on were included, which is not the case for the phonemic discrimination. For the phonemic discrimination part of the lexical decision task, both versions of the word were included as one singular entity in the analysis, thus creating *item pairs* of which one was incorrect and the other correct. This means that each *pair* is included in the analysis, not each individual item.

After the incorrect, timed-out and practise answers were deleted, several files were created for analysis. The original data had all answers from all participants in a single file, so for analysis, a new file was created which contained the percentages of correct and incorrect answers per participant. The d-prime scores for all participants were calculated to check for a response bias in their answers. The d-prime scores for the filler items (i.e., the items measuring lexical decision ability) were used to assess if pupils needed to be excluded due to response bias. The reason only the filler items are used for this purpose is because the target items are meant to be more difficult, and lower d-prime scores (i.e., indicatory of a response bias) are thus more likely to be found and not necessarily a reason to exclude pupils. A response bias on a difficult task indicates the strategy that a participant used for answering items they did not know the answer to, and can therefore still be informative. The filler items are, in this regard, used as control items as these are all real words that the participants could be familiar with, since this part of the task is not supposed to be difficult enough that the participants should be guessing for a large number of trials. Participants with d-prime scores below 0 on the filler items are thus considered to have guessed too often even during the relatively simpler part of the task, and will be excluded so that they do not skew the results.

The d-prime is preferred to calculating accuracy in scores because it allows us to measure how well participants perform while also controlling for response bias/guessing. On basis of the d-primes of the filler items, one Year 4 pupil was excluded. While their accuracy score is not sensitive enough to indicate a problem with their answers, the score below 0 on the d-prime indicates that this participant was not sensitive to the items when answering: i.e., they may have always answered *correct* or *incorrect*. By doing so, they may have an accuracy score close to 50% (i.e., chance level) which does not actually reflect the pupils' ability to recognise words. With this pupil removed from the dataset, all participants that are included are shown in Figure 2. Since only one pupil was excluded, it suggests this part of the task is not so difficult for the pupils, as they did not guess. This indicates that using the filler items as a control to check whether pupils were guessing during the test or not was a valid way with which to decide whether pupils needed to be excluded. Since the filler items and the target items all come from the same lexical decision task, this also demonstrates that the pupils did not guess on the phonemic discrimination items either despite lower d-prime scores. Instead, their lower d-prime scores indicate that the pupils in the current study experience more difficulty with discriminating phonemes than they do with recognising existing words and rejecting non-words.

Figure 2
Plot of d-prime and accuracy scores



4.3.3 Statistical analysis

All analyses were conducted in IBM SPSS Statistics 27. The dependent variable was either the score as measured with accuracy or with the d-prime for each pupil. The independent variables were each pupils' school year (Year 1, 2 or 4) and whether they followed the TTO-track or not (TTO or non-TTO). The scores on the PPVTs served as a covariate. However, as not all pupils participated in the PPVT tests (either the Dutch, the English, or both), the PPVTs could only be included in a smaller sample of the scores. With the use of two-way ANOVAs, the interactions between TTO and school year were calculated. The effect of the independent variables (school year and TTO) and covariates (both PPVTs, when included) on the pupils' d-prime scores and accuracy scores on both phonemic discrimination and lexical decision ability was calculated. The findings of these analyses can be found in Sections 5.1 and 5.2.

A separate analysis, conducted with a one-way repeated measures ANOVA, was done to analyse how pupils' scores on phonemic discrimination ability were affected by the manner of articulation of the item pairs. The only independent variables included in this analysis was manner of articulation The items were subdivided in 1) vowels (/æ/-/ɛ/), 2) fricatives (/s/-/z/ and /f/-/v/) and 3) plosives (/p/-/b/ and /t/-/d/). It was hypothesised that item pairs including the /æ/-/ɛ/ vowels would be harder to phonemically discriminate in comparison to the consonants (fricatives and plosives) as pupils may not be familiar with this contrast due to the lack of phonemic contrast between these sounds in Dutch. Additionally, a two-way ANOVA was conducted on the d-prime and accuracy scores on each separate manner of articulation with school year and TTO as independent variables. This analysis was conducted in order to assess whether school year and TTO affected the scores on a certain manner of articulation. For example, TTO pupils or pupils in later years may perform better on fricatives or plosives, but may still find vowels too hard to discriminate. The findings of the analyses described in this paragraph can be found in Section 5.3.

5 Results

5.1 Lexical decision ability

5.1.1 The effect of school year and TTO on lexical decision ability

To begin with, the filler items measuring lexical decision ability (as a measure of vocabulary size) were investigated. Before conducting a series of analyses of variance, the d-prime and accuracy scores were inspected. The scatterplot in Figure 3 illustrates the scores on the d-prime and accuracy for all pupils (divided by what school year they were in and whether they were in the TTO-track or not). As becomes visible in Figure 3, all pupils had a d-prime above 0 and an accuracy above 50% (as explained in Section 4.3.2, the only participant who had a d-prime score below 0 was removed from the study). The lowest scores were from pupils in their first year, as well as one non-TTO pupil in their second year. The highest scores were from fourth-year TTO pupils. In Table 2, the descriptive statistics are provided to give an overview of how well pupils in each group performed. The averages of the d-prime and accuracy scores seemed to improve with each year and were generally higher for the TTO pupils.

The analyses described in the next paragraphs will assess whether these observed trends were statistically significantly different and to what degree TTO and school year were able to affect pupils' scores on lexical decision ability.

Figure 3

Pupils' scores on lexical decision

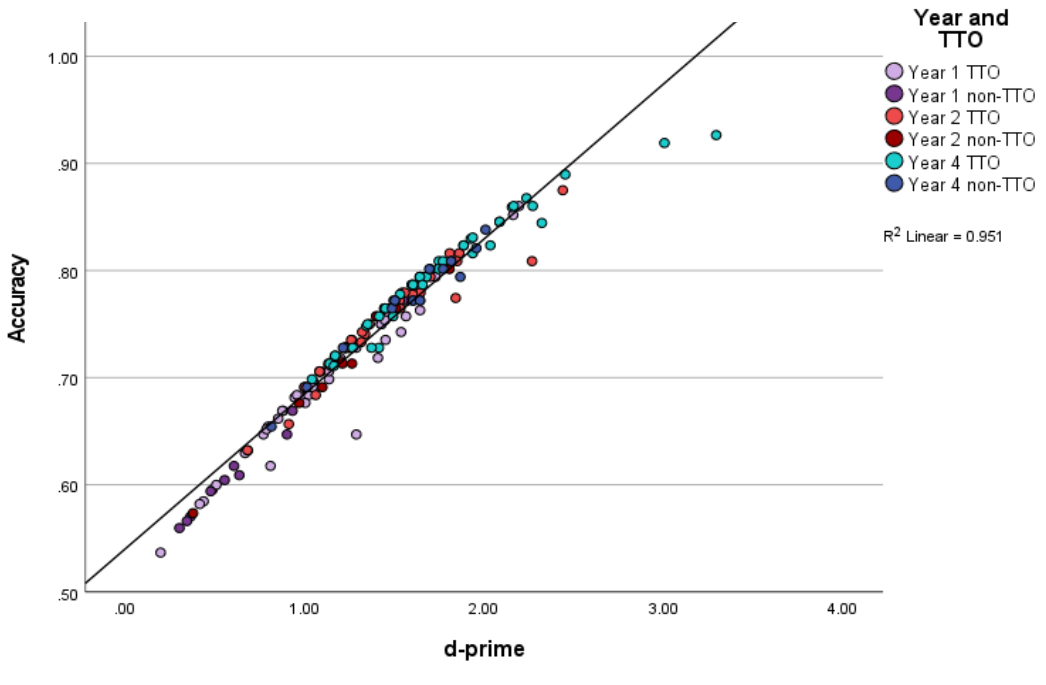


Table 2
Means and standard deviations (between parentheses) of the pupils' scores on lexical decision ability per group, split by school year and TTO

TTO			Non-TTO		
Year	D-prime	Accuracy	Year	D-prime	Accuracy
Year 1 (n = 52)	1.166 (0.421)	.705 (.066)	Year 1 (n = 14)	0.815 (0.440)	.649 (.077)
Year 2 (n = 40)	1.416 (0.353)	.750 (.048)	Year 2 (n = 10)	1.272 (0.402)	.725 (.068)
Year 4 (n = 40)	1.769 (0.475)	.799 (.080)	Year 4 (n = 15)	1.57 (0.034)	.771 (.048)

D-prime

A two-way ANOVA with the dependent variable d-prime and the independent variables TTO and school year was conducted first (see Table 3). This model explained just over 30% of the variance (adjusted R^2 = .305). Tukey post-hoc comparisons (see Table 4) revealed there was a significant difference between all the school years (p < .001 between all years). Year 4 scored highest (M = 1.71), followed by Year 2 (M = 1.39) and lastly Year 1 (M = 1.09). School year had a large effect size (η_p^2 = .26). The analysis revealed a main effect of school year (p < .001): pupils in later school years scored significantly different on the filler items in comparison to pupils in lower years (i.e., they had a better vocabulary and awareness of real words versus non-words) regardless of whether they followed TTO. Taken together, this confirms the trend observed earlier that higher school years perform better in comparison to lower years.

Furthermore, the analyses revealed a main effect of TTO (p = .003). Pupils in the TTO track (M = 1.42) performed significantly better on lexical decision ability than non-TTO pupils (M = 1.22). There was a slightly under-medium effect of TTO (η_p^2 = .054). However, there was no interaction between school year and TTO (p = .534, η_p^2 = .008). This indicates that pupils who were in the TTO track started off with better lexical decision ability than non-TTO pupils, but did not improve significantly stronger in their scores over time in comparison to pupils in the monolingual track.

Table 3

Output of between-subjects ANOVA with d-prime as dependent variable

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	13.620	5	2.724	15.925	.000	.325
Intercept	208.793	1	208.793	1220.620	.000	.881
Year	10.074	2	5.037	29.448	.000	.263
TTO	1.607	1	1.607	9.395	.003	.054
Year * TTO	.216	2	.108	.630	.534	.008
Error	28.224	165	.171			
Total	367.234	171				
Corrected Total	41.844	170				

Note. R Squared = .325 (Adjusted R Squared = .305)

Table 4

Post-hoc comparisons (Tukey) by school year with d-prime as dependent variable

Contrast	Mean Difference	95% Confidence Interval		*SE*	*p*
		Lower Bound	Upper Bound		
Year 1 vs Year 2	-0.300	-0.484	-0.117	0.078	<.001
Year 1 vs Year 4	-0.623	-0.801	-0.444	0.076	<.001
Year 2 vs Year 4	-0.322	-0.513	-0.131	0.081	<.001

Accuracy

A second two-way between-subjects ANOVA was conducted. It included the same independent variables as the ANOVA just reported (school year and TTO), but the dependent variable was the accuracy score instead of d-prime. School year and TTO remained the independent variables. This model for the analysis depicted in Table 5 explained nearly 35% of the variance (adjusted R^2 = .347), which is a larger percentage than the model with the d-prime as dependent variable managed to explain. Just like the analysis with the d-prime as dependent variable, the analysis with the dependent variable accuracy revealed a main effect of school year (p < .001). The effect of the school year in this model is large (η_p^2 = .305). Tukey post-hoc comparisons (see Table 6) revealed a significant difference between all the school years, regardless of TTO. The pupils of Year 4 scored highest (M = .79), followed by the pupils of Year 2 (M = .75) and lastly Year 1 (M = .69).

The analysis also revealed a main effect of TTO (p < .001). The effect was medium-sized (η_p^2 = .062). Overall, TTO-pupils scored significantly higher at lexical decision (M = .75) compared to non-TTO pupils (M = .72). There was again no interaction effect between TTO and school year, demonstrating that pupils in the TTO track did not improve significantly more over time in comparison to the non-TTO pupils. The implications are the same as for the analysis reported before. This current analysis with accuracy as dependent variable yielded similar results as the analysis with d-prime as dependent variable.

Table 5
Output of between-subjects ANOVA with accuracy as dependent variable

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	.337	5	.067	19.068	.000	.366
Intercept	62.952	1	62.952	17801.113	.000	.991
Year	.256	2	.128	36.154	.000	.305
TTO	.039	1	.039	10.918	.001	.062
Year * TTO	.006	2	.003	.870	.421	.010
Error	.584	165	.004			
Total	94.512	171				
Corrected Total	.921	170				

Note. R Squared = .366 (Adjusted R Squared = .347)

Table 6
Post-hoc comparisons (Tukey) by school year with accuracy as dependent variable

Contrast	Mean Difference	95% Confidence Interval		*SE*	*p*
		Lower Bound	Upper Bound		
Year 1 vs Year 2	-.052	-.078	-.026	.011	<.001
Year 1 vs Year 4	-.098	-.124	-.072	.011	<.001
Year 2 vs Year 4	-.046	.018	.073	.011	<.001

5.1.2 The effect of vocabulary size on lexical decision ability

Effect of TTO and school year on PPVT

Next, the relationship between lexical decision ability and vocabulary size in Dutch and English, as measured through the Dutch and English PPVTs, was investigated. Note that not all pupils participated in the PPVT tests: in some cases, they only completed one of the two or neither. These pupils were not included in this analysis with the vocabulary size as covariate. That is, in comparison to the prior analyses, the dataset used for the present analysis includes two additional covariates (the scores on the Dutch and English PPVTs), but is based on a smaller sample, i.e., it includes *some* of the pupils but not all. Of the total number of 171 pupils who completed the lexical decision task, a total of 113 participants completed the Dutch PPVT, and a total of 89 participants completed the English PPVT. A total of 60 pupils completed both the Dutch and English PPVTs.

In Table 7, the correlations between the Dutch and English PPVT scores and d-prime and accuracy scores on lexical decision are shown. The PPVT Dutch significantly correlated with the PPVT in English ($r(57) = .66$) as well as with the d-prime ($r(110) = .35$) and accuracy ($r(110) = .37$). The PPVT English correlated with d-prime ($r(87) = .41$) and accuracy ($r(87) = .41$), which were slightly stronger correlations in comparison to the PPVT Dutch. D-prime and accuracy strongly correlated with each other ($r(169) = .98$).

Table 7
Correlations between the PPVT scores (Dutch and English), d-prime and Accuracy

		PPVT Dutch	PPVT English	d-prime	Accuracy
PPVT Dutch	Pearson Correlation	1	.659**	.349**	.369**
	p (2-tailed)		.000	.000	.000
	N	112	59	112	112
PPVT English	Pearson Correlation	.659**	1	.410**	.413**
	p (2-tailed)	.000		.000	.000
	N	59	89	89	89
d-prime	Pearson Correlation	.349**	.410**	1	.975**
	p (2-tailed)	.000	.000		.000
	N	112	89	171	171
Accuracy	Pearson Correlation	.369**	.413**	.975**	1
	p (2-tailed)	.000	.000	.000	
	N	112	89	171	171

Note. ** Correlation is significant at the 0.01 level (2-tailed).

A two-way ANOVA was conducted with the dependent PPVT Dutch score (see Table 8) and the dependent PPVT English score (see Table 9) in order to assess whether the pupils' school year and TTO-track significantly affected scores on the PPVTs. It may be expected that TTO affected scores on the English PPVT but not on the Dutch PPVT, as this was the native language of both the TTO and non-TTO groups and should not matter for vocabulary size in Dutch. However, it is likely that school year affected both the English and Dutch PPVT scores, as older pupils will continue to learn new words in both languages no matter what track they are in.

The model with the Dutch PPVT as the dependent variable (with school year and TTO as independent variables) was able to explain over 55% (adjusted R^2 = .556) of the variance. A significant main effect of school year (p < .001) and not TTO (p = .060) was found, as predicted. There was a large effect size of school year (η_p^2 = .297). There was no interaction between school year and TTO (p = .159, η_p^2 = .018), indicating that TTO pupils did not improve to a significantly better degree on the Dutch PPVT than non-TTO pupils over time.

Table 8

Output of between-subjects ANOVA with PPVT Dutch as dependent variable

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	14150.055	4	3537.514	35.749	.000	.572
Intercept	695168.798	1	695168.798	7025.076	.000	.985
Year	4469.415	2	2234.707	22.583	.000	.297
TTO	356.719	1	356.719	3.605	.060	.033
Year * TTO	199.365	1	199.365	2.015	.159	.018
Error	10588.222	107	98.955			
Total	2518109.000	112				
Corrected Total	24738.277	111				

Note. R Squared = .572 (Adjusted R Squared = .556)

A two-way ANOVA with the PPVT in English as the dependent variable (with as independent variables school year and TTO) was conducted (see Table 9). The model was able to explain 36.7% of the variance (adjusted R^2 = .366). A significant main effect with a large effect size was found for school year (p < .001, η_p^2 = .20), indicating that pupils in later years outperformed pupils in lower years regardless of education type. Additionally, a main effect with a large effect size was found for TTO (p < .001, η_p^2 = .149), as was previously hypothesised. English vocabulary size was thus affected by TTO, but Dutch vocabulary size was not. Similarly to previous analyses, no interaction between school year and TTO was found (p = .313, η_p^2 = .012), indicating that pupils in the TTO-track did not have an advantage in improving their lexical decision ability over time in comparison to non-TTO pupils. Compared to the model with Dutch PPVT as the dependent, the analysis with the PPVT English had smaller effect sizes (adjusted R^2 = .366 for the English PPVT versus adjusted R^2 = .556 for the Dutch PPVT). This indicates that the independent variables (school year and TTO) more strongly affected performance on the Dutch PPVT than on the English PPVT.

Table 9

Output of between-subjects ANOVA with PPVT English as dependent variable

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	26024.549	4	6506.137	13.712	.000	.395
Intercept	1155939.878	1	1155939.878	2436.285	.000	.967
Year	10180.329	2	5090.164	10.728	.000	.203
TTO	6978.664	1	6978.664	14.708	.000	.149
Year * TTO	488.001	1	488.001	1.029	.313	.012
Error	39855.339	84	474.468			
Total	1920034.000	89				
Corrected Total	65879.888	88				

Note. R Squared = .395 (Adjusted R Squared = .366)

D-prime

A two-way between-subjects ANOVA was conducted (see Table 10). This ANOVA included the d-prime scores on lexical decision ability as its dependent variable and the Dutch and English PPVT scores as covariates. The independent variables included were school year and TTO. The model was able to explain 33.2% of the variance (adjusted $R^2 = .332$). There was a main effect of school year with a large effect size ($p = .027$, $\eta_p{}^2 = .130$). Neither the Dutch PPVT ($p = .525$, $\eta_p{}^2 = .008$) nor the English PPVT ($p = 759$, $\eta_p{}^2 = .002$) significantly influenced the pupils' lexical decision ability. No main effect of TTO was found, indicating that pupils who followed the TTO-track did not score significantly different compared to non-TTO pupils. There was no interaction between TTO and school year ($p = .828$, $\eta_p{}^2 = .001$).

Table 10

Output of between-subjects ANOVA with d-prime as dependent variable

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	6.395	6	1.066	5.798	.000	.401
Intercept	.156	1	.156	.848	.361	.016
PPVT Dutch	.075	1	.075	.409	.525	.008
PPVT English	.017	1	.017	.095	.759	.002
Year	1.429	2	.715	3.887	.027	.130
TTO	.184	1	.184	1.002	.322	.019
Year * TTO	.009	1	.009	.048	.828	.001
Error	9.559	52	.184			
Total	110.010	59				
Corrected Total	15.954	58				

Note. R Squared = .401 (Adjusted R Squared = .332)

Accuracy

Another two-way between-subjects ANOVA was conducted (see Table 11). It included the same independent variables as the ANOVA just reported (school year and TTO) and the same covariates (the English and Dutch PPVT scores), but the dependent variable was accuracy instead of d-prime. This model explained 33.5% of the variance (adjusted R^2 = .335), only a marginally larger percentage than the analysis with the d-prime as dependent variable. The only main effect found was school year (p = .014), which had a large effect size (η_p^2 = .152). TTO failed to be a statistically significant factor. Neither the Dutch PPVT (p = .592, η_p^2 =.006) nor the English PPVT (p = .745, η_p^2 = .002) affected pupils' lexical decision ability. Again, there was no interaction between TTO and school year (p = 698, η_p^2 = .003). The current analysis with accuracy as the dependent variable yielded similar results as the prior analysis with d-prime as the dependent variable.

Table 11

Output of between-subjects ANOVA with accuracy as dependent variable

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	.160	6	.027	5.866	.000	.404
Intercept	.101	1	.101	22.161	.000	.299
PPVT Dutch	.001	1	.001	.291	.592	.006
PPVT English	.000	1	.000	.107	.745	.002
Year	.042	2	.021	4.647	.014	.152
TTO	.004	1	.004	.907	.345	.017
Year * TTO	.001	1	.001	.152	.698	.003
Error	.237	52	.005			
Total	31.175	59				
Corrected Total	.397	58				

Note. R Squared = .404 (Adjusted R Squared = .335)

5.1.3 Filler item analysis

The filler items were used to measure lexical decision ability (representing vocabulary size) in the current task. D-prime cannot be calculated for the filler items as they were not paired (like the target items were): it is impossible to calculate both hits and false alarms for these items (which are required to calculate the d-prime score) as an item could *only* have a hit/miss or a false alarm/correct negative. For the filler items, the difficulty is thus measured with the accuracy percentage instead.

The average accuracy scores of the filler items can be seen in Table 12, divided in non-words and real words. Pupils had the largest issues with identifying real words rather than non-words. The participants only had accuracy scores of 50% or less for two non-words and 14 real words. Interestingly, most of the non-word items have a rather high average accuracy score, with most of the items falling between 60% and 89% accuracy. Interestingly enough, most of the items with an average accuracy score between 90% and 100% are real words.

Table 12
Average accuracy scores on lexical decision ability items

Accuracy Percentage	Real words	Non-words
10-19	gleam; meek; jeer	
20-29	drowse; yearn; chap	
30-39	scheme; shrill; vase; crook; spouse	hoke; parf
40-49	frown; crumb	
50-59	lash; weep; youth; drake; stir; mole; growl	swut; shorp
60-69	sake; crate; glaze	clum; wath; hidge; strawn; mome; sporf; fluss; shoul; firp; hube
70-79	dread; chief; curl; rage; blaze; cruel; crane; soak; south; boil; flush; steam	mosh; chert; stirl; trorse; drune; jark; crale; crale; shipe; lir; hean; spudge; gleathe; pute; dawsh
80-89	bath; grape; crown; foam; dish; shade	stroil; shrit; brile; squayle; vike; sprull; brutt; lon; drile; rouch; clil; strit; neve; plorn; trave; yearl; kiff; deadge; brong; feuth; trif; chadge; vabe; spetch; vodge; brear; murp; boin; parve; breen; dreeve; grish
90-100	sneeze; judge; dare; full; thin; pipe; fark; duck; suck; search; scare; screen; chair; fine; blood; guide; mouse; skill; church; nice; game; shine; ship; street; snake; flight; touch	fraf; quirze; sman; meach; froop; fub; waph; noik

5.2 Phonemic discrimination ability

5.2.1 The effect of TTO and school year on phonemic discrimination ability

Figure 4 shows the scatterplot of the pupils' ability to discriminate phonemes measured by d-prime and accuracy. These scores were calculated based on the target words in the test: real words that were either phonetically manipulated (near-words) or not. As can be seen in Figure 4, the scores of the pupils were more varied than they were for the lexical decision ability items (i.e., the filler items). Phonemic discrimination was, as was expected, relatively challenging for the pupils to do correctly: the scores were generally lower and more scattered than scores on lexical decision ability. For instance, the highest score belonged to a TTO pupil in Year 4, but the second- and third-best pupils were first-years: one who followed the TTO track and one who did not. Some Year 4 pupils also fell in the lower ranges of scores, including pupils who were in the TTO track. Pupils with a d-prime below 0 (a total of 30 pupils) can be generally considered to have done poorly at phonemic discrimination in this task.

Furthermore, in Figure 4 and Table 13, it can be noticed that there was a trend for TTO pupils to generally perform better than non-TTO pupils, and for pupils in higher years to have better scores than pupils in lower years (except for the non-TTO pupils in Year 4, who scored lower than Year 2). Statistical analyses will assess whether these differences were significant and whether TTO and school year affected pupils' ability to discriminate phonemes.

Figure 4

Pupils' scores on phonemic discrimination

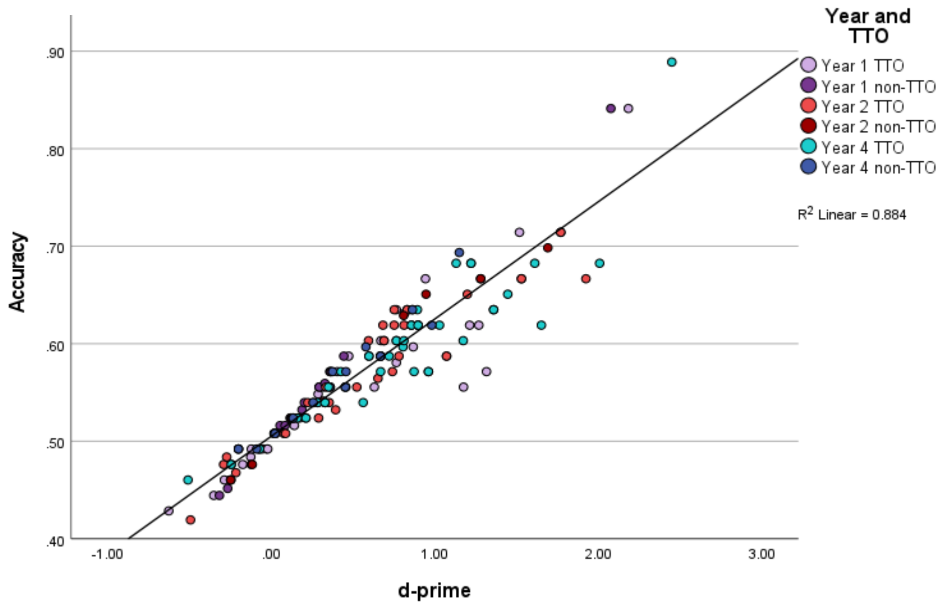


Table 13

Means and standard deviations (between parentheses) of the pupils' scores on phonemic discrimination ability per group, split by school year and TTO

TTO			Non-TTO		
Year	D-prime	Accuracy	Year	D-prime	Accuracy
Year 1 (*n* = 52)	0.415 (0.583)	.555 (.077)	Year 1 (*n* = 14)	0.383 (0.606)	.564 (.098)
Year 2 (*n* = 40)	0.597 (0.562)	.572 (.064)	Year 2 (*n* = 10)	0.556 (0.696)	.572 (.094)
Year 4 (*n* = 40)	0.713 (0.612)	.587 (.076)	Year 4 (*n* = 15)	0.408 (0.422)	.564 (.059)

D-prime

A two-way ANOVA was conducted with the dependent variable d-prime (see Table 14). The independent variables were TTO and school year. The analysis revealed that there were no significant main effects. School year did not significantly influence pupils' performance on phonemic discrimination (p = .312, η_p^2 = .014). There was also no significant effect of TTO (p = .243, η_p^2 = .008). No significant interaction effect between school year and TTO was found (p = .479, η_p^2 = .009). This model explained a mere 1.8% of the variance (adjusted R^2 = .018) and was thus not able to explain variance well. These results indicate that the trends observed earlier were not significant: the pupils' d-prime scores measuring their phonemic discrimination ability were similar across the three school years and across the TTO and non-TTO track.

Table 14

Output of between-subjects ANOVA with d-prime as dependent variable

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	2.737	5	.547	1.616	.158	.047
Intercept	30.738	1	30.738	90.764	.000	.355
Year	.795	2	.398	1.174	.312	.014
TTO	.465	1	.465	1.373	.243	.008
Year * TTO	.501	2	.251	.740	.479	.009
Error	55.878	165	.339			
Total	107.111	171				
Corrected Total	58.615	170				

Note. R Squared = .047 (Adjusted R Squared = .018)

Accuracy

Another two-way ANOVA was conducted with the dependent variable accuracy (see Table 15). The independent variables were school year and TTO. The analysis revealed that there were no significant main effects. The pupils' school year was not a significant factor in this model (p = .583, $\eta_p{}^2$ = .007). TTO was not a significant main effect (p = .746, $\eta_p{}^2$ = .001), indicating there was no considerable difference between the scores of TTO pupils and non-TTO pupils. Again, there was no significant interaction effect between school year and TTO (p = .604, $\eta_p{}^2$ = .006). This model fit the data poorly and the predictors were not at all able to explain the variance (adjusted R^2 = -.005). Pupils across all groups should score similarly. This model with the accuracy scores as dependent variable yields similar results as the model with the d-prime scores as dependent variable.

Table 15
Output of between-subjects ANOVA with accuracy as dependent variable

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	.024	5	.005	.841	.522	.025
Intercept	37.943	1	37.943	6689.745	.000	.976
Year	.006	2	.003	.541	.583	.007
TTO	.001	1	.001	.106	.746	.001
Year * TTO	.006	2	.003	.505	.604	.006
Error	.936	165	.006			
Total	56.342	171				
Corrected Total	.960	170				

Note. R Squared = .025 (Adjusted R Squared = -.005)

5.2.2 The effect of vocabulary size on phonemic discrimination ability

TTO and school year on PPVT

In Section 5.1.2, it was analysed whether TTO and school year were able to influence pupils' scores on the English and Dutch PPVT scores. As this analysis was not affected by either the scores on lexical decision ability or phonemic discrimination ability, such an analysis will not be repeated in this current section. As explained in Section 5.1.2, the dataset including the PPVT scores of Dutch and English as covariates was smaller. Only 60 pupils of the total of 171 who completed the lexical decision task were included. In the next paragraphs, the effect of the vocabulary size in Dutch and English (as measured by the PPVT scores) on phonemic discrimination will be analysed.

In Table 16, the correlation between the PPVT scores in English and Dutch and the phonemic discrimination scores, measured by d-prime and accuracy, are shown. The scores on the PPVT Dutch significantly correlated with the scores on the English PPVT (r(58) = .65) and with the d-prime (r(111) = .22) but surprisingly enough not with accuracy (r(111) = .10). The PPVT English correlated with neither d-prime (r(87) = .16) nor with accuracy (r(87) = .08). Accuracy and d-prime correlated strongly (r(169) = .94).

Table 16

Correlations between PPVT scores (Dutch and English), d-prime and accuracy

		PPVT Dutch	PPVT English	d-prime	Accuracy
PPVT Dutch	Pearson Correlation	1	.648**	.223*	.101
	p (2-tailed)		.000	.018	.285
	N	113	60	113	113
PPVT English	Pearson Correlation	.648**	1	.156	.084
	p (2-tailed)	.000		.145	.434
	N	60	89	89	89
d-prime	Pearson Correlation	.223*	.156	1	.940**
	p (2-tailed)	.018	.145		.000
	N	113	89	171	171
Accuracy	Pearson Correlation	.101	.084	.940**	1
	p (2-tailed)	.285	.434	.000	
	N	113	89	171	171

Note. **. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

D-prime

A two-way between-subjects ANOVA was conducted (see Table 17). The dependent variable was the d-prime and the independent variables were TTO and school year. The covariates PPVT English and PPVT Dutch were included as well. The analysis revealed that there was a lack of significant main effects. Neither TTO (p = .188, η_p^2 = .033) nor school year (p = .897, η_p^2 = .004) significantly influenced pupils' ability to discriminate phonemes. No significant interaction effect between school year and TTO was found (p = .435, η_p^2 = .012). This model fit the data poorly; the independent variables were not able to explain the variance (adjusted R^2 = -.037).

Table 17

Output of between-subjects ANOVA with d-prime as dependent variable

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	1.425	6	.237	.653	.688	.069
Intercept	.021	1	.021	.058	.810	.001
PPVT Dutch	.286	1	.286	.787	.379	.015
PPVT English	.237	1	.237	.651	.423	.012
Year	.079	2	.039	.109	.897	.004
TTO	.648	1	.648	1.782	.188	.033
Year * TTO	.225	1	.225	.618	.435	.012
Error	19.276	53	.364			
Total	39.223	60				
Corrected Total	20.701	59				

Note. R Squared = .069 (Adjusted R Squared = -.037)

Accuracy

Another two-way ANOVA was conducted (see Table 18). Compared to the previous analysis, this model included the same independent variables (school year and TTO) and the same covariates (PPVT Dutch and PPVT English), but the dependent variable was accuracy instead of d-prime. The model revealed there were no significant main effects. There was no effect of TTO (p = .373, $\eta_p{}^2$ = .015) and no effect of school year (p = .944, $\eta_p{}^2$ = .002). There was no significant interaction effect between school year and TTO (p = .508, $\eta_p{}^2$ = .008). This model fit the data poorly; the independent variables were not at all able to explain the variance (adjusted R^2 = -.080). The current analysis with the dependent accuracy yielded similar results as the preceding analysis with the dependent d-prime.

Table 18

ANOVA with accuracy as dependent variable

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	.009	6	.002	.268	.949	.029
Intercept	.076	1	.076	12.980	.001	.197
PPVT Dutch	.000	1	.000	.069	.794	.001
PPVT English	.005	1	.005	.934	.338	.017
Year	.001	2	.000	.057	.944	.002
TTO	.0005	1	.005	.807	.373	.015
Year * TTO	.003	1	.003	.444	.508	.008
Error	.311	53	.006			
Total	19.807	60				
Corrected Total	.320	59				

Note. R Squared = .029 (Adjusted R Squared = -.080)

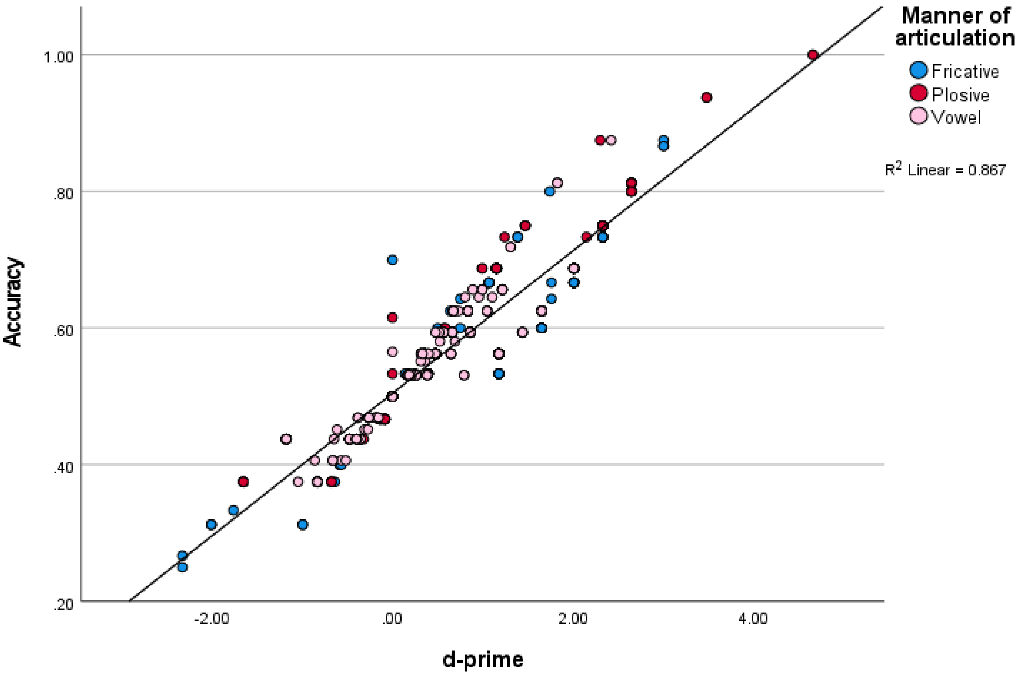
5.3 Manner of articulation

5.3.1 The effect of manner of articulation on phonemic discrimination ability

In this section, the effect of the three tested manners of articulation (plosives, fricatives and vowels) on pupils' phonemic discrimination ability will be discussed. The pupils' scores on each of the three manners of articulation were calculated in order to assess how well they do on each of these. The scatterplot in Figure 5 illustrates the scores on both accuracy and d-prime of all pupils per manner of articulation (i.e., all pupils have three dots: one for their score on fricatives, one for their score on plosives, and one for their score on vowels). The highest scores that were reached (i.e., the items that are most easily phonemically discriminated) were plosives. The lowest scores on the scatterplot refer all to items with fricatives. Judging by Figure 5, there seems to be a trend for plosives to be easier to discriminate for the pupils, and for fricatives to be more difficult. This would not be in line with the hypothesis that the vowels were the hardest to discriminate. The analyses in the next few paragraphs will assess whether this trend was statistically significant or not.

Figure 5

Pupils' scores on phonemic discrimination by manner of articulation



D-prime

First of all, we inspected the correlations (Pearson's correlation coefficient) between the d-prime scores on the three different manners of articulation. Table 19 reveals that the pupils' d-prime scores on all manners of articulation were positively correlated ($p < .001$). That is, a pupil who performed well on one type of manner of articulation was more likely to have higher scores on the other items as well.

Table 19

Correlation between pupils' d-prime scores on the manners of articulation

		Fricative	Plosive	Vowel
Fricative	Pearson Correlation	1	.285**	.289**
	p (2-tailed)		.000	.000
	N	172	172	172
Plosive	Pearson Correlation	.285**	1	.282**
	p (2-tailed)	.000		.000
	N	172	172	172
Vowel	Pearson Correlation	.289**	.282**	1
	p (2-tailed)	.000	.000	
	N	172	172	172

Note. Correlation is significant at the 0.01 level (2-tailed).

A one-way repeated measures ANOVA (see Table 20) with the d-prime scores as the dependent and manner of articulation (with three levels: vowels, fricatives and plosives) as independent variables was performed. This was done in order to investigate whether the pupils' phonemic discrimination abilities were different for vowel, fricative and plosive contrasts. The analysis revealed a significant effect of manner of articulation on d-prime scores (F (1, 3) = 60.691, $p < .001$), as well as a large effect size for manner of articulation ($\eta_p{}^2 = .27$). Post-hoc comparisons were conducted with a paired-samples t-test. These post-hoc comparisons (see Table 21) revealed that d-prime scores were significantly different between plosives and vowels ($p < .001$) as well as plosives and fricatives ($p < .001$) but not between fricatives and vowels ($p = .111$). The pupils were best able to discriminate plosives ($M = 1.12$) and able to discriminate fricatives ($M = 0.55$) and vowels ($M = 0.40$) to the same degree.

Table 20

Output of the one-way repeated measures ANOVA with d-prime as dependent variable

Source of variation	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Between groups	509.686	1	509.585	364.057	.000	.682
Within groups	109.964	3	36.655	60.691	.000	.263
Error	308.015	510	.604			
Total	927.665	514				

Table 21
Post-hoc comparisons (paired-samples t-test) by manner of articulation with d-prime as dependent variable

Contrast	Mean Difference	95% Confidence Interval		*SE*	*p*
		Lower Bound	Upper Bound		
Fricative vs plosive	-0.566	-0.772	-0.361	0.104	<.001
Plosive vs vowel	0.714	-0.559	0.870	0.079	<.001
Vowel vs fricative	-0.148	-0.331	0.034	0.093	.111

Accuracy

Next, the correlations (Pearson's correlation coefficient) between the accuracy scores on the three different manners of articulation were investigated. Table 22 reveals that, just like the d-prime scores, the accuracy scores on all three types of manners of articulation were significantly and positively correlated. Not that the correlations were stronger than for those measured with d-prime. Pupils who attained high scores on one type of manner of articulation thus typically performed well on the other item types too.

Table 22
Correlation between pupils' accuracy scores on the manners of articulation

		Fricative	Plosive	Vowel
Fricative	Pearson Correlation	1	$.328^{**}$	$.320^{**}$
	p (2-tailed)		.000	.000
	N	172	172	172
Plosive	Pearson Correlation	$.328^{**}$	1	$.337^{**}$
	p (2-tailed)	.000		.000
	N	172	172	172
Vowel	Pearson Correlation	$.320^{**}$	$.337^{**}$	1
	p (2-tailed)	.000	.000	
	N	172	172	172

Note. Correlation is significant at the 0.01 level (2-tailed).

A second one-way repeated measures ANOVA (see Table 23) was conducted with the dependent variable accuracy instead of d-prime. The independent variable was manner of articulation (vowels, fricatives, and plosives). The analysis revealed a significant effect of manner of articulation on accuracy (F (1, 3) = 183.470, p < .001). Manner of articulation had a large effect size (η_p^2 = .519). Post-hoc comparisons (see Table 24) revealed a statistical difference between plosives and fricatives (p < .001) as well as between plosives and vowels (p < .001), but no statistical difference between vowels and fricatives was found (p = .835). The pupils were best able to discriminate plosives (M = .63), followed by fricatives (M = .55) and vowels (M = .55). The findings for the model with the dependent variable d-prime scores were similar as for the current findings for the model which used the accuracy scores as the dependent variable.

Table 23

Output of the one-way repeated measures ANOVA with d-prime as dependent variable

Source of variation	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Between groups	260.380	1	260.380	13095.846	.000	.987
Within groups	4.119	3	1.373	183.470	.000	.519
Error	3.816	510	.007			
Total	268.315	514				

Table 24

Post-hoc comparisons (paired-samples t-test) by manner of articulation with accuracy as dependent variable

Contrast	Mean Difference	95% Confidence Interval		*SE*	*p*
		Lower Bound	Upper Bound		
Fricative vs plosive	-.076	-.097	-.054	.011	<.001
Plosive vs vowel	.078	.060	.095	.009	<.001
Vowel vs fricative	-.002	-.022	.018	.010	.835

5.3.2 The effect of school year and TTO on phonemic discrimination ability per manner of articulation

The analysis on manner of articulation on pupil-level revealed that manner of articulation does have a large effect on the pupils' scores on phonemic discrimination, both when measured by d-prime and average accuracy. The following analyses will assess whether the accuracy and d-prime scores per manner of articulation are affected by TTO and school year. It is possible that school year and TTO influenced the types of manner of articulation in a different way. In that case, perhaps TTO and school year may have a significant effect on the scores on fricatives or plosives, for example, but not on vowels. That is, such a difference could be caused by the vowel discrimination being hypothetically more difficult to acquire: this could lead to additional language exposure having no effect on pupils' improvement on phonemic discrimination of vowels. This section will first report the analysis for the d-prime scores and then for the accuracy scores. In Figure 6, 7 and 8, the scatterplots for the d-prime and accuracy pupils' scores on each manner of articulation type can be seen, divided by TTO and school year. No clear trend immediately emerges from the scatterplots, so no statistical differences between TTO and school year on phonemic discrimination were expected to be found.

Figure 6

Pupils' scores on phonemic discrimination of plosives

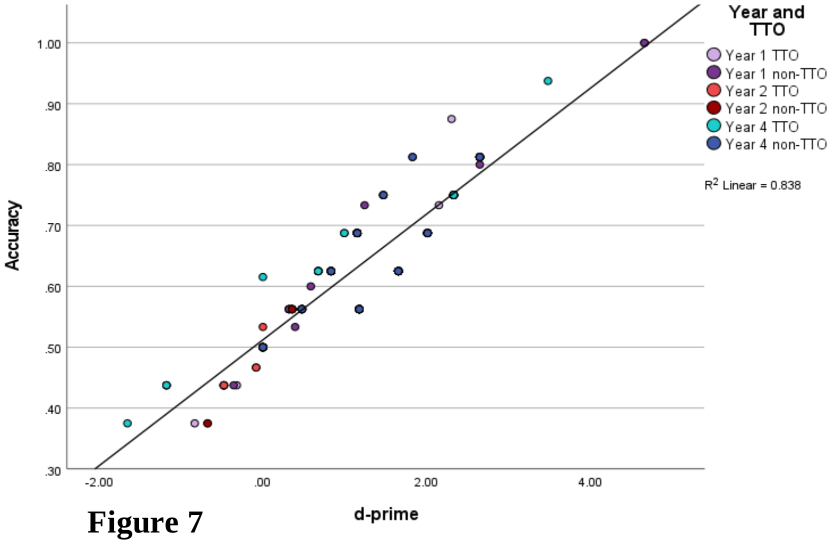


Figure 7

Pupils' scores on phonemic discrimination of fricatives

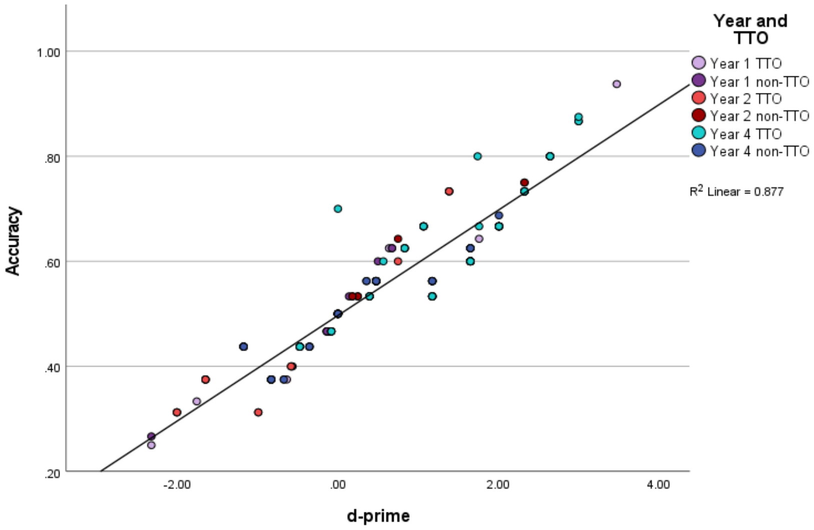
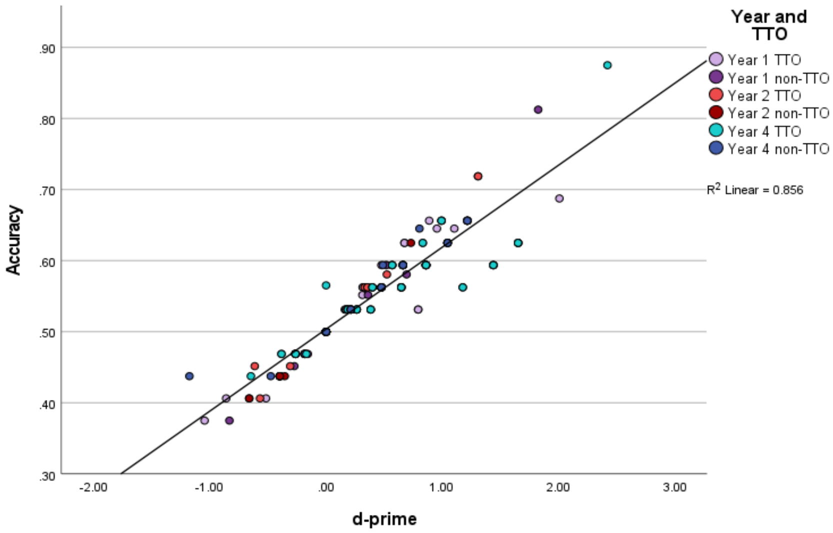


Figure 8

Pupils' scores on phonemic discrimination of vowels



D-prime

Plosives

A two-way ANOVA with the d-prime scores on plosive items as dependent variable and the independent variables TTO and school year was conducted (see Table 25). The model was not able to explain the variance (adjusted R^2 = -.010). The analysis revealed no main effects of TTO (p = .956, η_p^2 < .001) nor of school year (p = .911, η_p^2 = .001). There also was no interaction between school year and TTO (p = .426, η_p^2 = .010). This indicates that pupils overall did not improve their skills to discriminate plosives over time, and that TTO pupils were not more skilled or improved to a significantly better degree than non-TTO pupils. In other words, all pupils – across TTO and across school years – performed similarly, which is visualised in Figure 9.

Table 25

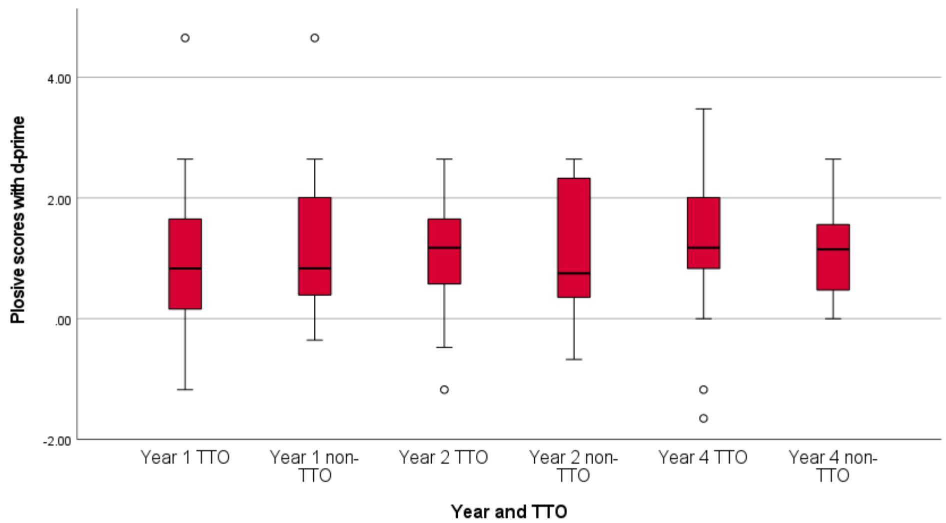
Output of the two-way ANOVA with dependent variable d-prime scores on plosives

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	3.440	5	.688	.651	.661	.019
Intercept	149.268	1	149.268	141.203	.000	.460
Year	.197	2	.098	.093	.911	.001
TTO	.003	1	.003	.003	.956	.000
Year * TTO	1.813	2	.907	.858	.426	.010
Error	175.481	166	1.057			
Total	392.404	172				
Corrected Total	178.921	171				

Note. R Squared = .019 (Adjusted R Squared = -.010)

Figure 9

Boxplot of d-prime scores on plosives per school year and TTO



Fricatives

A two-way ANOVA was conducted with d-prime scores on items with fricatives as dependent and the independent variables TTO and school year (see Table 26). This model explained only 3.9% of the variance (adjusted $R^2 = .039$). The analysis revealed there was no main effect of TTO ($p = .223$, $\eta_p^2 = .009$); however, there was a significant main effect of school year with a medium effect size ($p = .036$, $\eta_p^2 = .039$). Although according to the descriptive statistics, the pupils in Year 2 performed best ($M = 0.73$), closely followed by Year 4 ($M = 0.73$) and with the first-year pupils performing worst ($M = 0.25$), these observed mean differences may not be significant. That is, Tukey's test for post-hoc (see Table 27) revealed no significant difference between school years, which was surprising given the significant main effect. Pupils in Year 2 and Year 4 may have scored similarly but possibly outperformed pupils in Year 1. This may indicate that the pupils' ability to phonemically discriminate fricatives has improved upon after the first year, but then pupils did not improve between Year 2 and Year 4. Since Tukey's test for post-hoc did not confirm a significant difference, however, we do not want to overstate the effect of school year on pupils' ability to discriminate fricatives. There was no interaction between school year and TTO ($p = .210$, $\eta_p^2 = .019$), meaning that the effect of school year on fricative discrimination was not further modulated by TTO. The scores of all groups are visualised in Figure 10.

Table 26

Output of the two-way ANOVA with dependent variable d-prime scores on fricatives

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	17.557	5	3.511	2.378	.041	.067
Intercept	30.203	1	30.203	20.457	.000	.110
Year	9.984	2	4.992	3.381	.036	.039
TTO	2.213	1	2.213	1.499	.223	.009
Year * TTO	4.656	2	2.328	1.577	.210	.019
Error	245.089	166	1.476			
Total	314.255	172				
Corrected Total	262.646	171				

Note. R Squared = .067 (Adjusted R Squared = .039)

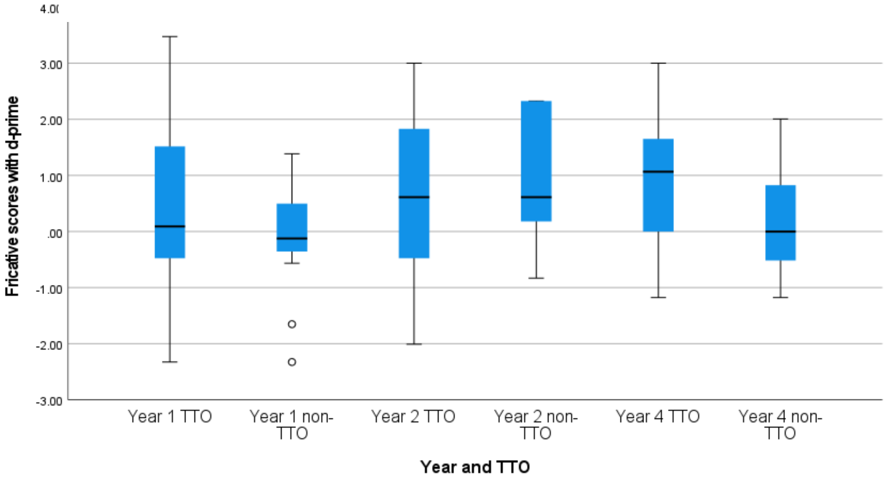
Table 27

Post-hoc comparisons (Tukey) by year with the dependent variable d-prime scores on fricatives

Contrast	Mean Difference	95% Confidence Interval		*SE*	*p*
		Lower Bound	Upper Bound		
Year 1 vs Year 2	-0.493	-1.031	0.046	0.228	.081
Year 1 vs Year 4	-0.480	-1.001	0.043	0.221	.079
Year 2 vs Year 4	0.013	-0.546	0.573	0.236	.998

Figure 10

Boxplot of d-prime scores on fricatives per school year and TTO



Vowels

Another two-way ANOVA was conducted with the accuracy scores on plosive items as dependent variable and the independent variables TTO and school year (see Table 28). This model was not able to explain the variance (adjusted R^2 = -.014). The analysis revealed that there were no main effects of TTO (p = .345, η_p^2 = .005) nor of school year (p = .661, η_p^2 = .005), as well as a lack of interaction between school year and TTO (p = .895, η_p^2 = .001) . This indicates that pupils did not improve the skills to discriminate vowels over time, and that TTO pupils were not more skilled or improved more strongly over time in comparison to non-TTO pupils. This similarity between all pupils, regardless of TTO or school year, is illustrated in the boxplot in Figure 11.

Table 28

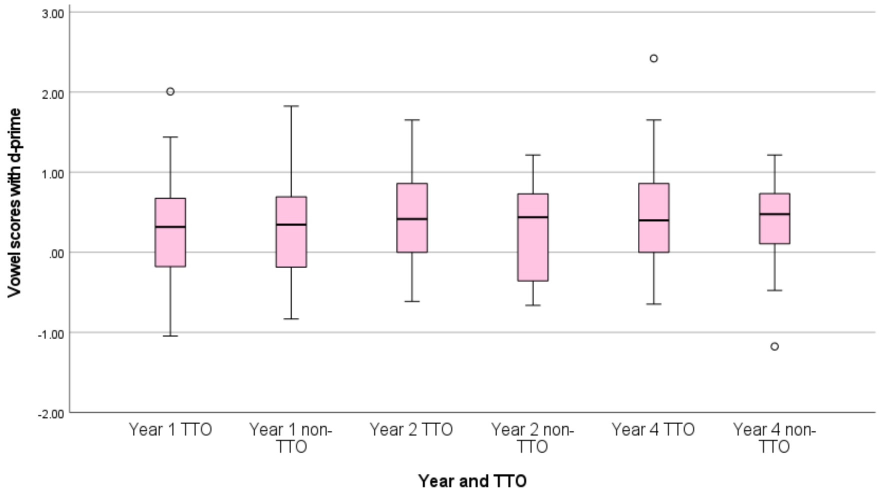
Output of the two-way ANOVA with dependent variable d-prime scores on vowels

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	1.014	5	.203	.532	.752	.016
Intercept	16.538	1	16.538	43.368	.000	.207
Year	.317	2	.158	.415	.661	.005
TTO	.343	1	.343	.898	.345	.005
Year * TTO	.085	2	.042	.111	.895	.001
Error	63.303	166	.381			
Total	91.774	172				
Corrected Total	64.318	171				

Note. R Squared = .016 (Adjusted R Squared = -.014)

Figure 11

Boxplot of d-prime scores on vowels per school year and TTO



Accuracy

Plosive

Another two-way ANOVA was conducted with accuracy scores on plosive items as the dependent variables and the independent variables TTO and school year (see Table 29). This model was not able to explain the variance (adjusted R^2 = -.010). The analysis revealed there were no main effects of TTO (p = 568, η_p^2 = .002) nor of school year (p = .698, η_p^2 = .004). There was no interaction effect between school year and TTO found (p = .915, η_p^2 = .011). The analysis using the d-prime scores on the plosives as a dependent yielded similar results as this current analysis. The findings indicate that pupils did not improve their skills to discriminate plosives over time, and that TTO pupils were not more skilled or and did not improve more strongly over time. All pupils performed similarly, as is illustrated in in Figure 12.

Table 29

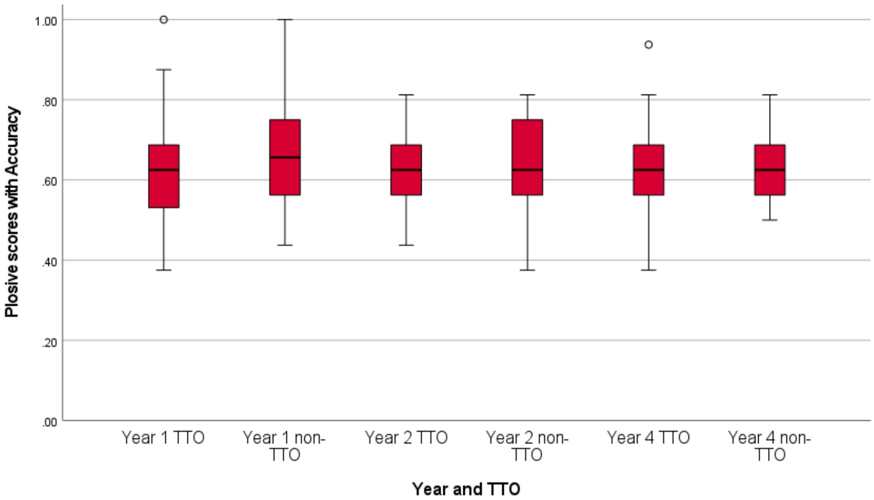
Output of two-way ANOVA with dependent variable accuracy scores on plosives

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	.046	5	.009	.677	.641	.020
Intercept	46.743	1	46.743	3459.432	.000	.954
Year	.010	2	.005	.360	.698	.004
TTO	.004	1	.004	.327	.568	.002
Year * TTO	.025	2	.012	.915	.403	.011
Error	2.243	166	.014			
Total	70.002	172				
Corrected Total	2.289	171				

Note. R Squared = .020 (Adjusted R Squared = -.010)

Figure 12

Boxplot of accuracy scores on plosives per school year and TTO



Fricative

A two-way ANOVA was conducted with the accuracy scores on fricative items as the dependent and the independent variables TTO and school year (see Table 30). This model was able to explain a small part of the variance (adjusted $R^2 = .035$) Unlike the analysis using the d-prime scores as the dependent variable (see above), this analysis found no main effect of school year ($p = .077$, $\eta_p^2 = .030$). No significant main effect of TTO was found ($p = .343$, $\eta_p^2 = .005$), and there was no significant interaction between school year and TTO ($p = .135$, $\eta_p^2 = .024$). Figure 13 shows the comparison of the pupils' scores per group.

Table 30

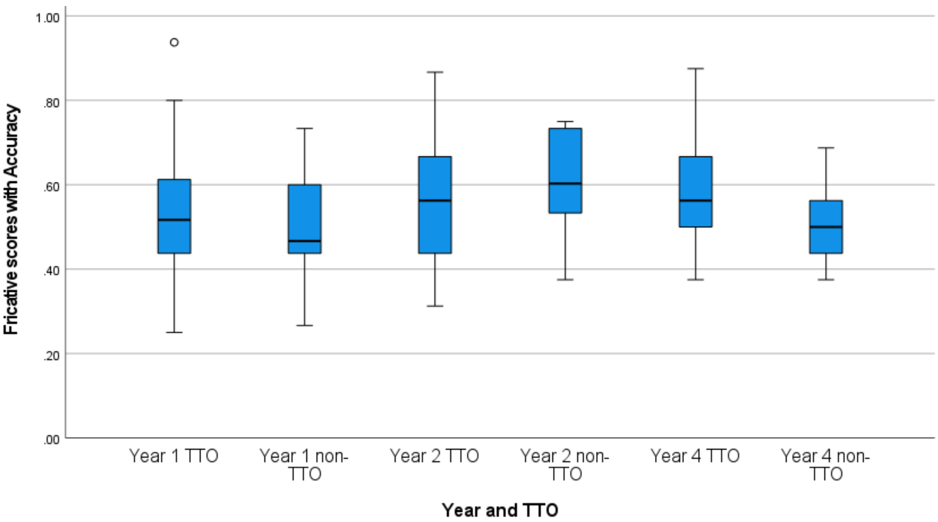
Output of two-way ANOVA with dependent variable accuracy scores on fricatives

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	.190	5	.038	2.238	.053	.063
Intercept	35.434	1	35.434	2085.527	.000	.926
Year	.089	2	.044	2.610	.077	.030
TTO	.015	1	.015	.903	.343	.005
Year * TTO	.069	2	.034	2.028	.135	.024
Error	2.820	166	.017			
Total	55.387	172				
Corrected Total	3.011	171				

Note. R Squared = .063 (Adjusted R Squared = .035)

Figure 13

Boxplot of accuracy scores on fricatives per school year and TTO



Vowel

Another two-way ANOVA was conducted with the accuracy scores on vowel item pairs as dependent variable and the independent variables TTO and school year (see Table 31). This model was not able to explain the variance (adjusted R^2 = -.020). The analysis revealed there are no main effects of TTO (p = .782, η_p^2 < .001) nor of school year (p = .606, η_p^2 = .006). There was no interaction between school year and TTO (p = .833, η_p^2 = .002). These findings are similar to the findings of the analysis using the d-prime scores as dependent. Pupils did not improve upon the ability to discriminate vowels over time and TTO pupils did not have better phonemic discrimination abilities, nor did they improve significantly stronger over time in comparison to non-TTO pupils. As can be seen in Figure 14, all pupils across all groups score similarly.

Table 31

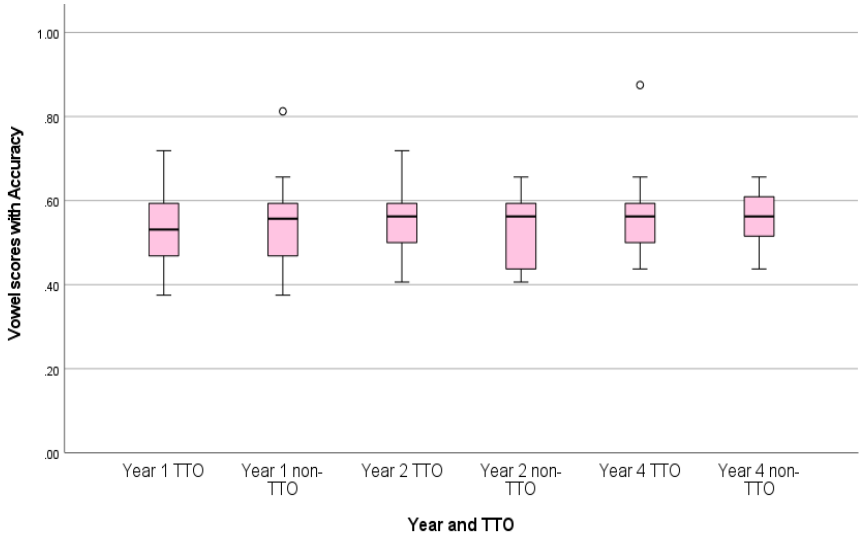
Output of two-way ANOVA with dependent variable accuracy scores on vowels

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Corrected Model	.010	5	.002	.341	.887	.010
Intercept	35.378	1	35.378	5938.433	.000	.973
Year	.006	2	.003	.502	.606	.006
TTO	.000	1	.000	.077	.782	.000
Year * TTO	.002	2	.001	.183	.833	.002
Error	.989	166	.006			
Total	52.983	172				
Corrected Total	.999	171				

Note. R Squared = .010 (Adjusted R Squared = -.020)

Figure 14

Boxplot of accuracy scores on vowels per school year and

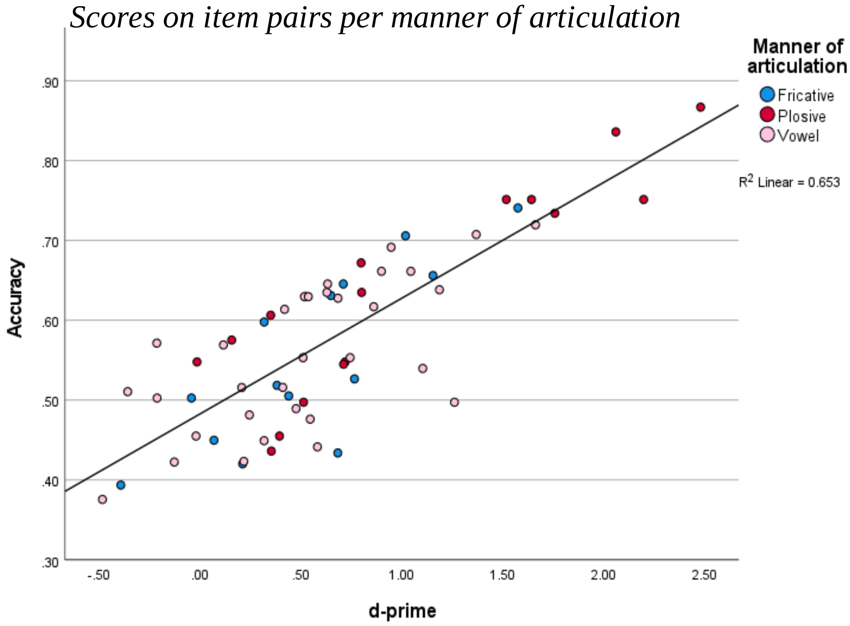


5.3.3 Target item analysis

The aim of this section is to look at the d-prime scores on individual item pairs (the target items measuring phonemic discrimination ability) and to assess why certain item pairs may have been more difficult or easy to discriminate in comparison to others. In Figure 15, the d-prime and accuracy scores of all item pairs can be found. The figure illustrates that there was a lot of variation between item pairs within the same manner of articulation. Certain item pairs with plosives were among the easiest, but other item pairs with the same manner of articulation were seemingly more difficult to perform well on. Item pairs containing fricatives and vowels seem to vary between d-prime scores of -0.5 and 2, and item pairs with plosives seem to vary between d-prime scores of 0 and 2.5.

Figure 15

Scores on item pairs per manner of articulation



The d-prime scores of all item pairs can be found in Table 32. The items with the lowest d-primes are the ones that pupils had the largest issues with. A total of nine item pairs had d-primes scores under 0, which indicates that participants often guessed on these item pairs. The most difficult item pair was “bench” with its near-word “banch” (an item pair in which the stem vowel [ɛ] was switched with [æ]). Six of the nine hardest item pairs include the vowel manipulation. The remaining three include two fricative manipulations from voiced to voiceless: [s] to [z] (“choice” and “choise”) and [f] to [v] (“dwarf” and “dwarve”). The last one is an item pair which includes the manipulation of a voiced plosive to a voiceless one: [b] to [p] (“globe” and “glope”). These findings seem to imply that pupils may struggle most to correctly accept and reject items with the vowel contrast in the lexical decision task, perhaps because the non-native contrast is difficult to perceive. The analysis of the role of manner of articulation (see Section 5.3.1) assessed whether there were statistical differences between the item pairs and how pupils performed at item pairs with a certain manipulation. In this analysis, we found that pupils generally performed best at discriminating real words and near words for which the manner of articulation investigated were the plosives, as there was a statistical difference between item pairs with vowels/fricatives and item pairs with plosives. Item pairs with fricatives and vowels were not statistically different from each other.

All item pairs with a d-prime above 0 are not considered particularly difficult for participants. The higher the d-prime score, the better pupils scored on a certain item pair. As can be seen in Table 32, some of the highest scores were on plosives, especially the /b/-/p/ item pairs. Three /b/-/p/ item pairs (“sheep”/“sheeb”, “cheap”/“cheab” and “ship”/“shib”) were seemingly the easiest pairs for which to discriminate phonemes considering the d-prime score of above 2. It can be seen that there is a wide spread of variation of the score for all manners of articulation. Even if plosives are considered easier to discriminate phonemically, there are still item pairs with plosives that are among the hardest trials (“globe”/“glope” has a d-prime of below 0). There are item pairs among the fricatives and vowels with d-prime scores above 1.5 (“move”/“moof” and “black”/“bleck”, respectively). Most of the /æ/-/ɛ/ item pairs have an average d-prime score between 0.5 and 1.

Table 32
Items pairs split by manner of articulation and d-prime scores

D-prime	Item pairs				
	Plosives		Fricatives		Vowels
	/d/-/t/	/b/-/p/	/v/-/f/	/z/-/s/	/æ/-/ɛ/
-0.5 to 0		globe / glope	dwarf / dwarve	choice / choise	bench / banch; quest / quast; thank / thenk; smell / small; pram / prem; chess / chass
0 to 0.5	glide / glite; smart / smard	tube / tupe; cube / cupe;	scarf / scarve; laugh / lauve; shave / shafe	phrase / phrace; news / newce	wealth / walth; gram / grem; chest / chast; breast / brast; press / prass; spank / spenk; sweat / swat; slam / slem
0.5 to 1	skirt / skird; blade / blate beard / beart spit / spid	rub / rup	stiff / stiv; dive / dife	nurse / nurze; voice / voise; cheese / cheece	bank / benk; blank / blenk; scratch / scretch; trap / trep; fresh / frash; swell / swall; jet / jat; breath / brath; lamp / lemp; desk / dask; death / dath; rank / renk
1 to 1.5			groove / groof	kiss / kiz	plank / plenk; dress / drass; splash / splesh; smash / smesh; span / spen
1.5 to 2	proud / prout; flight / flied	sharp / sharb	move / moof		black / bleck
2 to 2.5		sheep / sheeb; cheap / cheab; ship / shib			

6. Discussion

The present study investigated whether bilingual education led to a better performance on phonemic discrimination and lexical decision ability (measuring vocabulary size) in the L2 English and how non-TTO and TTO pupils improved over time on these two language areas. I will first restate the results and make some other remarks that I feel are important. Then I will discuss the answers to the research questions and any additional information important to these questions in the order in which the questions were posed. The discussion section has been divided into three parts: the first deals with lexical decision ability (RQ1 and RQ2), the second discusses phonemic discrimination ability (RQ3, RQ4 and RQ5), and the third deals with the suitability of the d-prime and accuracy scores in analysing the data (RQ6).

The results revealed that TTO and school year significantly influenced the pupils' scores on lexical decision ability (representing vocabulary knowledge), with TTO pupils outperforming non-TTO pupils, and older pupils outperforming pupils in lower school years. Both the Dutch and English PPVT tests significantly correlated with the pupils' scores on lexical decision ability, but did not constitute significant covariates in an analysis on lexical decision scores. This means that while there is a correlation between Dutch and English vocabulary size on the pupils' performance on lexical decision ability, we cannot claim that a larger Dutch and English vocabulary size as measured by the PPVTs significantly affected the pupils' vocabulary size as measured by the lexical decision task.

Neither school year nor TTO were found to have significantly affected phonemic discrimination ability, and neither PPVT tests constituted significant covariates. This indicates that vocabulary size in neither English nor Dutch, as measured by the PPVTs, significantly influenced the pupils' phonemic discrimination ability in their L2 English. Surprisingly, only the Dutch PPVT, and not the English PPVT, correlated significantly with phonemic discrimination ability, but only when measured with the d-prime. The pupils' scores on phonemic discrimination appeared to depend on the manner of articulation of the tested contrasts, with items that contained plosives seemingly being easiest to discriminate for pupils, followed by fricatives and vowels. Pupils scored similarly on items with fricatives and items with vowels, but when looking at individual items they scored lowest on, the item pairs with vowels were among those. Against the expectations, TTO pupils did not perform better on any given manner of articulation type than non-TTO pupils, and pupils overall did not improve their phonemic discrimination skills over time, except perhaps for fricative discrimination.

Additionally, it is important to note that pupils generally did better on the filler items (measuring vocabulary size) than they did on the target items (measuring phonemic discrimination ability). No statistical analysis was run on the difference between these items, as this contrast was not the aim of this current thesis, but pupils scored relatively poorly on the target items in comparison to the filler items, with the d-prime scores and accuracy scores on the filler items being consistently higher. This was expected beforehand, as the filler items measure a language area that we expected the pupils to score better on compared to phonemic discrimination, and the phonemic contrasts chosen for this study are known to be difficult to perceive for native speakers of Dutch (Broersma & Cutler, 2008).

Discussion on lexical decision ability

RQ1: The effect of TTO and school year on lexical decision ability

I will first discuss the first research question, which poses the question whether school year and TTO affected lexical decision and if there was an interaction between school year and TTO. It was found that both TTO and school year affected lexical decision ability (as a proxy for vocabulary size) significantly: TTO pupils outperformed non-TTO pupils and pupils overall improved over time. Our finding that TTO pupils outperformed non-TTO pupils on lexical decision ability seems, at first, a sign that TTO plays a beneficial role in pupils' vocabulary size. However, the lack of an interaction effect of TTO and school year revealed that, against our expectations, TTO did not enable pupils to improve to a significantly higher degree. That is, TTO pupils did not improve more strongly over time than non-TTO pupils. It seems that they did not necessarily acquire proportionally more vocabulary throughout high school than non-TTO pupils did, even though they had a larger vocabulary size to start with. A possible way to explain the current study's findings is that TTO pupils might already have had a higher proficiency in certain language skills, including vocabulary knowledge. This is also a possible reason for why they may have been attracted to join a TTO track in the first place. Therefore, we cannot conclude from the findings of the present study that TTO played a beneficial role in the TTO pupils' performance. Another factor (e.g., pupils' motivation to learn language or TTO pupils have a better language aptitude than non-TTO pupils) that we did not investigate might be responsible for TTO pupils' larger vocabulary size.

RQ2: The relation of vocabulary size and lexical decision ability

The second research question was concerned with the degree to which lexical decision ability and the two vocabulary tests (the Dutch and English PPVTs) were related. The pupils' scores on lexical decision ability were found to correlate with both the Dutch PPVT and the English PPVT, with a slightly stronger correlation for the English one. As the lexical decision task was only in English, this last finding is not surprising. Unexpectedly, neither PPVT was a significant covariate for lexical decision ability, indicating that performance on either of the PPVTs does not significantly influence performance on lexical decision ability (although the performances are *related*, as revealed by the correlation test). Although the English PPVT and the lexical decision ability part of the task supposedly measure the same knowledge (receptive vocabulary size in English), the tasks may still be rather different internally from each other, entailing that pupils may have scored differently on both. While the analysis without the PPVTs as covariates revealed a main effect of TTO on lexical decision ability (see RQ1), the analysis that included the PPVTs as covariates does not. The dataset which included the PPVTs as covariates was smaller as well, due to some pupils being excluded because they did not complete all tests, and perhaps some nuance was missed that was included in the analyses of the larger dataset that excluded the PPVT scores.

Additionally, another analysis was conducted with the PPVTs as the dependent variable, rather than the covariate, to see whether school year and TTO influenced the performance on both PPVT tests. A main effect was found for both TTO and school year for the English PPVT and only school year for the Dutch PPVT. This was expected, as TTO pupils would have had more exposure to English and were therefore expected to have a larger vocabulary, but the scores in Dutch were presumed to be similar between the TTO and non-TTO groups. This is also in line with the findings for vocabulary size as measured by the lexical decision ability (which was *only* in English), as TTO and school year were both significant factors for these (see above). In this analysis, no interaction effect was found between school year and TTO, again affirming that TTO pupils did *not* improve significantly stronger over time in comparison to non-TTO pupils. These findings are in agreement with other studies on the advantage of bilingual education on pupils' vocabulary size, which found that TTO pupils outperformed those who did not have bilingual education, but did not necessarily prove that they improved more strongly over time (Admiraal et al., 2006; Alonso et al., 2008; Verspoor et al., 2015). Once again, this leads us to wonder whether it is TTO that has a beneficial role in the acquisition of further vocabulary skills during high school or whether other factors (e.g., motivation to learn target language or a higher language aptitude in comparison to non-TTO pupils) cause TTO pupils to have a larger vocabulary size than non-TTO pupils. To summarise the answer to this research question: lexical decision ability and vocabulary size in both Dutch and English are correlated, but vocabulary size in either language does not influence pupils' scores on lexical decision ability.

On the difficulty of the lexical decision ability items

Although no specific research question was formed to discuss specific items, Section 5.1.3 discussed whether certain trials were particularly hard or easy for the pupils to correctly recognise as a real word or to correctly reject as a non-word. This section aims to discuss the possible reason for these difficulties in more depth.

Mostly, it seemed that pupils scored worst on existing words and incorrectly identified them as non-words. A possible cause may be that pupils, when in doubt, chose to identify any words they were not familiar with as a non-word. When they did this for non-words, they answered correctly. However, if they chose this approach for real words, their answers would be incorrect and the overall accuracy of that particular item would decrease. Pupils were thus more likely to have a miss (an incorrect answer on an existing word) than a false alarm (an incorrect answer on a non-word). The real words they fail to recognise are most likely words they encounter infrequently, whereas the non-words they are identifying as existing words may have a resemblance to words they do know or sound familiar to the pupils. However, pupils generally scored very well on real words, with 71% of the real word items having an average accuracy score above 60% (and 39% of the real word items having an accuracy percentage of above 90%). It seems that the pupils are familiar with certain words and correctly recognise them most of the time, but have low scores on any words that they are less familiar with. Generally, the pupils did well on recognising non-words as well, with 94% of the non-words having an average accuracy score of above 60%.

Discussion on phonemic discrimination ability

RQ3: The effect of TTO and school year on phonemic discrimination ability

The third research question regarded the effect of school year and TTO on phonemic discrimination and whether there was an interaction effect between these two variables. According to the analysis of the target items of the lexical decision task, TTO and school year had no effect on phonemic discrimination ability. Pupils across both language tracks and across the three school years appeared to score similarly on phonemic discrimination ability. These findings suggest that language track and school year did not affect the pupils' ability to discriminate phonemes. The fact that the models were not able to explain the variance and fit the data extremely poorly further demonstrates that TTO and school year did not influence pupils' phonemic discrimination ability.

A possible explanation is that TTO does not offer enough or qualitatively sufficient language exposure for TTO pupils to improve upon their phonemic discrimination ability or, alternatively, that exposure alone is not enough. If the additional language exposure that TTO pupils receive was indeed enough for them to improve their ability to discriminate phonemes, an improvement over time should have been found in the data. That is, the amount of language exposure all pupils have had increased over time for both TTO and non-TTO pupils (although TTO pupils had the benefit of receiving more language exposure in general). Since we have found in this study that the older pupils did not outperform pupils in lower years, it becomes clear that language experience alone was not capable of teaching pupils to discriminate these phonemic contrasts. Perhaps the quality of language exposure is not sufficient to allow pupils to improve their ability to discriminate the tested phonemes. Other studies found that quality of exposure can lead to differences in the acquisition process of phonetic skills (Paradis, 2011; Uchihara et al., 2022). However, in the current study, we do not have any information about the quality of language exposure. Moreover, since all pupils were taught by the same teachers, quality of language exposure was not a factor that could be tested in this current study. Alternatively, supporters of bilingual education sometimes argue that a lack of improvement in a language skill may simply be due to a ceiling effect. That is, after a certain level of proficiency has been reached, it becomes impossible to improve further (Verspoor et al., 2015). However, this explanation is unlikely to be the case in the present study considering the relatively poor average performance on phoneme discrimination. Considering the fact that the pupils in this study were at the start of their secondary school career, the second-year and fourth-year pupils should have shown some improvement in the phoneme discrimination ability if the additional exposure and language learning in TTO were to be advantageous in this facet of language acquisition. The fact that no such improvement was found makes it clear that additional language exposure does not necessarily lead to an improvement of pupils' ability to distinguish between phonemic contrasts.

The results of this current study suggest that in order to improve upon certain phonemic discrimination ability in a foreign language, the additional language exposure of TTO was not sufficient. It would be interesting to investigate if pupils' phonemic discrimination abilities would improve after being *explicitly* taught the differences as was the case for the participants in Lacabex & Gallardo-del-Puerto (2020)'s study on schwa discrimination. Such an experiment would allow for us to control for the quality and quantity of the exposure. Agustín Llach (2017) theorises after her study on lexical production that pupils who follow bilingual education may only gain an advantage when there is a combination of implicit language exposure with a more explicit approach. Possibly, in a language area that is considered more challenging than vocabulary acquisition, such as phonemic discrimination or other phonetic skills, this explicit approach may be especially necessary. Little research has been conducted on the advantages of bilingual education on the acquisition of phonetic perception. It is well-known that phonetic details are the hardest to acquire in a new language, as phonemic perception and production are often largely influenced by the native language which can impede the acquisition of non-native phonemic categories (Baker & Trofimovich, 2005; Bosch et al., 2000). However, how formal education influences the acquisition of these trickier aspects of a foreign language appears to have received little attention. One might expect the additional exposure of bilingual education to allow pupils to be more familiar with non-native phonemic contrasts, but it may be that this aspect of language learning is too difficult to grasp without proper training, or that it requires much more exposure to the target language. Perhaps schools should pay more attention to phonetic information in language class and offer more explicit instruction. Another study could focus on whether such an approach might lead to a difference in pupils' ability to discriminate non-native phonemic contrasts over time. If both TTO and non-TTO pupils receive explicit instruction on phonemic differences in the target language, we could explore whether any additional language exposure that TTO pupils receive leads to a stronger improvement afterwards in comparison to non-TTO pupils.

Another possibility for both TTO and non-TTO pupils' low scores and lack of improvement on phonemic discrimination of these contrasts is that they these skills are harder to learn after a certain age. Section 2.2.2 discusses critical age and age effects on phonemic acquisition, and how certain studies have found that after age 12, there is a drastic decline in learners' ability to acquire certain language skills (Abu-Rabia & Kehat, 2004). There is a chance that pupils were too old to improve these phonemic discrimination skills and to become more proficient than they were before the time of this study. However, other studies did find that phonological ability improves after exposure. Moreover, an existence of age effects does not necessarily imply that pupils will find it *impossible* to improve; they just indicate that there are different mechanisms in place for phonological learning which may make it harder than for an adult learner than it is for a young learner. Therefore, it may be so that age effects affected pupils' phonemic discrimination ability, but they are necessarily the most important factor.

Additionally, a possible explanation is that the low scores on phonemic discrimination did not stem from pupils' inability to actually *discriminate* the phonemes. That is, perhaps the pupils were able to hear the phonemic differences, but did not reject inaccurate pronunciations because they were used to accepting these variations. After all, it has been noted how prevalent accented speech is even in highly fluent speakers (Wolfswinkler & Reinisch, 2016), which is likely to have caused Dutch pupils to be exposed to inaccurate pronunciations of these phonemes quite often. The phonemes that were tested in this study were specifically selected because of the difficulty they cause Dutch native speakers: the pupils in this study are likely to have been exposed to non-native pronunciations and may have learnt to accept this speech variation. This is a good thing, too, as otherwise speech comprehension would decrease and communication would suffer. In order to maintain successful conversations, Dutch pupils are encouraged to accept these inaccurate pronunciations and are less likely to reject words based on a variant phoneme, although a native speaker might when listening to L2-accented speech. The hypothesis that pupils are able to discriminate these phonemes could be tested in a future study in which the pupils are being asked to reject incorrect *pronunciations* instead of asked to reject non-existent words. Such a study would enable us to assess whether pupils may have been aware of the incorrect pronunciation in this study but accepted a near-word as being real because they have encountered a similar pronunciation before. In the present study, we cannot be certain of the strategy – if there was any – that the pupils applied. Some pupils may have achieved low scores on phonemic discrimination ability because they tended to accept words with incorrect phonemic variations as real words, despite being able to successfully discriminate the phonemes. Others may have followed the design as intended (i.e., rejecting incorrect pronunciations). To summarise the results of this research question: TTO and school year did not significantly affect pupils' ability to discriminate phonemes in the target language, and we should consider other factors to explain phonemic discrimination ability.

RQ4: The relation of vocabulary size and phonemic discrimination ability

The fourth research question investigated the relationship between the Dutch and English PPVTs and phonemic discrimination ability. The correlation test revealed that there was no significant correlation between the English PPVT and phonemic discrimination. There was, however, a significant correlation between the Dutch PPVT and the pupils' scores on phonemic discrimination ability, though only when the scores were measured with d-prime. Considering that phonemic discrimination ability was measured using only English words, this finding is surprising. Perhaps it is some measure of native language ability that allows pupils to be more attuned to phonemic differences, even when such differences belong to a foreign language. However, since the only language skill measured in Dutch was vocabulary size, we cannot be certain why Dutch vocabulary, and not English vocabulary, correlated positively with phonemic discrimination ability. Future research could investigate the potential role of the L1 on phonemic discrimination ability in the L2. Neither the Dutch nor English PPVT tests were found to be significant covariates when analysing phonemic discrimination ability: this means that neither Dutch nor English vocabulary size significantly affected pupils' phonemic discrimination ability. Dutch vocabulary size may be correlated to phonemic discrimination ability but did not affect it. This finding is not wholly surprising, considering the PPVT scores did not significantly affect for the lexical decision ability either, which were correlated to a far higher degree. Neither TTO nor school year were significant factors, similarly to the findings of the analysis without the PPVT scores as covariates (see Research question 3). There was no interaction effect between TTO and school year. All pupils across years and language tracks score similarly on phonemic discrimination. To conclude, the inclusion of the PPVTs did not affect the findings on significant factors on pupils' phonemic discrimination ability, nor did the PPVTs themselves significantly affect phonemic discrimination ability.

RQ5: The effect of manner of articulation on phonemic discrimination ability

The fifth research question examined whether there were any differences at the level of the pupils' phonemic discrimination ability between the three manners of articulation tested: the fricatives, plosives or vowels. The type of manner of articulation did seem to have a big impact on how well pupils performed on differentiating real words from near-words (with manipulated phonemes). Plosives were found to be relatively easier to discriminate for the pupils: they were significantly different from the scores on fricatives and vowels. Pupils performed similarly on item pairs containing fricatives and vowels. There were no statistical differences found between the pupils' scores on fricatives and vowels. These findings are not in line with the hypothesis that vowels would be hardest to discriminate for the pupils. The low scores on the item-pairs that included the vowels /æ/-/ɛ/ were expected beforehand, since this contrast is not found in the pupils' native language and it is well-known that this contrast is hard to distinguish for native Dutch speakers (Broersma & Cutler, 2008). However, the phonemic contrasts between the voiced and unvoiced phonemes of fricatives (respectively, /v/-/f/ and /z/-/s/) do exist in Dutch, and were therefore expected to be easier for the pupils in comparison to the non-native phoneme contrast of the vowels. This hypothesis does not seem to be valid in the present study, as pupils perform equally well on the fricatives and the vowels. Possibly, the ongoing trend in Dutch of the devoicing of the fricatives (De Schryver et al., 2013; Van de Velde & Van Hout, 2001) may play a role in this difficulty in discriminating fricatives, because voiced-unvoiced contrasts in fricatives are becoming less relevant to Dutch speakers. Over time, no significant difference was found regarding how TTO and non-TTO pupils improve on any of the different manners of articulation, or any statistical evidence that they improve at all. As for the general development (independent of language track) of phonemic discrimination ability over time, school year was only found to significantly affected the fricative scores, reflecting that the first-years did significantly worse than the second- and fourth-years. However, this effect was only found when the dependent variable was the d-prime score, and Tukey's post-hoc analysis did not confirm this trend. A future study could more deeply investigate the effect that different manners of articulation have on pupils' phonetic discrimination ability in order to analyse whether there truly is a statistical difference between the manners. In sum, pupils were best able to differentiate between item pairs containing unvoiced-voiced plosive contrasts. Neither TTO pupils nor non-TTO pupils improved over time depending on manner of articulation, except perhaps on the discrimination of fricatives, and TTO pupils did not outperform non-TTO pupils.

On the difficulty of the phonemic discrimination ability items

Similarly to the discussion on filler items, the target items measuring phonemic discrimination ability were also analysed in Section 5.3.3 to assess if certain item pairs were considered more difficult or particularly easy. Certain item pairs had a d-prime below 0, which indicates that pupils did badly on these trials (i.e., they were often unable to correctly accept a real word or to correctly reject as a near-word). On the other hand, other item pairs had high d-prime scores (of above 2) indicating that pupils were able to phonemically discriminate these words well. This section aims to discuss why certain item pairs may be more difficult or easy to discriminate for pupils.

Presumably, word frequency is an important factor. A word that occurs more frequently in a language is one that is more familiar to a speaker, and that familiarity might help to correctly accept/reject a certain pronunciation of the word. Additionally, perhaps a word with a low d-prime score may have been manipulated in such a way that the incorrect near-word still seems like a real word (although perhaps another one than intended by this study). For example, the "smell/small" manipulation (in which "smell" is the real word; the real word "small" is pronounced [smɔːl], whereas "small" in this study refers to the manipulation of the /ɛ/ vowel to become /æ/ in which "small" is a near-word with the pronunciation [smæl]). However, it may be possible that a word like "small" [smɔːl] was also activated as a possible meaning for this near-word. Additional phonological neighbours (i.e., the number of words that only have one different phoneme than the target word) do affect word processing. That is, words with more phonological neighbours are slower to be recognised and more mistakes are made in word recognition during auditory processing (Ziegler et al., 2003). However, a lack of phonological neighbours may also have caused difficulties. A word like "choice", for example, has fewer phonological neighbours and it is likely only the intended target word was activated during the task both when the real word and the near word were tested. This may be an explanation for why the item pair "choice"/"choise" is among the lowest d-prime scores: the pupils may have *always* recognised it as "choice" and thus made many errors when the near-word was tested. My expectation is that a combination of word frequency and phonological neighbourhood density has led to the differences in d-prime scores between item pairs with the same phoneme manipulations.

Furthermore, a possible explanation of the d-prime scores on some item pairs is that a manipulation in one direction may be harder to perceive than a manipulation in another direction. For example, the words in which the /b/-/p/ phonemes have been swapped have different d-prime scores dependent on the direction of the phoneme manipulation. All the item pairs in this category from -0.5 to 1 are item pairs in which the word-final /b/ was swapped for word-final /p/. However, the item pairs with d-prime scores above 1.5 have been manipulated with the word-final /p/ becoming /b/. If not for lack of time, it would have been interesting to conduct a statistical analysis with an additional factor that considered the direction of the phoneme manipulations. It seems that pupils found it easier to recognise that words that ought to end with a voiceless consonant (/p/) are incorrect when they are pronounced with a voiced consonant (/b/). Possibly, this has to do with their L1. In Dutch, word-final consonants are never voiced: Dutch native speakers tend to devoice these voiced consonants in English because of this aspect of their native language (i.e., it would not be unexpected to hear a native Dutch speaker pronounce “globe” [gləʊb] as “glope” [gləʊp]). However, it is unlikely that a native Dutch speaker would insert a voiced consonant in place of a voiceless one (i.e., it is unexpected to hear a native Dutch speaker pronounce “sharp” [ʃɑːp] as “sharb” [ʃɑːb]). This pattern of low d-prime scores depending on manipulation seems to mostly hold true for the /b/-/p/ item pairs and is less obvious in all other item pairs. The two highest /d/-/t/ items contain one of each manipulations, and the fricative item pair with the highest score (“move”/“moof”) is a voiced-to-voiceless manipulation (i.e., the opposite of what might have expected to find if we had hypothesised pupils to score better on item pairs of which the real word has a word-final voiceless consonant). Especially the /æ/-/ɛ/ manipulation might be different in comparison to the consonant manipulations, as this discrimination may be harder for pupils to perceive in the first place. There are both /æ/-to-/ɛ/ and /ɛ/-to-/æ/ manipulations among the item pairs with a d-prime score of -0.5 to 0. Among the item pairs with a d-prime above 1, however, only one out of a total of six items contain an /ɛ/-to-/æ/ manipulation. Perhaps this manipulation is considered more difficult to perceive correctly. This might be something for a future study to investigate.

Discussion on d-prime and accuracy scores

RQ6: The suitability of d-prime versus accuracy scores in analysing the data

The sixth and final research question, which was more of a methodological nature, was concerned with the suitability of the d-prime scores versus the accuracy scores in analysing the lexical decision task data. Not a great many differences were found between the results of the analyses using the d-prime or accuracy scores, respectively, most of the time. The effect sizes and percentage of variance explained were most often somewhat bigger for the d-prime but there were occasions where the accuracy scores seemed to fit the statistical model better. There were two occasions of the analysis with the d-prime that found a significant effect that did not occur in the same analysis with the accuracy as the dependent variable. The first was the analysis of the fricative items in order, assessing whether TTO and school year affected pupils’ ability to phonemically discriminate specific types of manner of articulation. The analysis with the d-prime as dependent variable revealed a main effect of school year that the analysis with the accuracy scores did not find. The second difference was in the correlation between the two PPVT tests and the scores on phonemic discrimination. The Dutch PPVT significantly correlated with the scores as measured by the d-primes but not accuracy. In general, all other findings were similar.

As has been shown in this thesis, the d-prime generally revealed stronger effect sizes and explained more variance. There were some exceptions in which the accuracy scores seemed to fit the statistical model better, but for most part the findings were largely similar. Moreover, the d-prime also allowed for a pre-statistical analysis to assess whether participants were scoring above chance and whether they were sensitive to distinguish between target and filler items. This allows us to observe response bias. Simply relying on accuracy would not have allowed the removal of the one participant who scored far below chance, but whose accuracy scores did not reflect the same inability to perform adequately on the test. Because people are less familiar with d-prime scores, the d-prime scores can be straightforward in the interpretation. However, I believe the d-prime is more suitable to use for this type of research in comparison to accuracy scores, and I recommend future studies to use d-prime scores as the dependent variable.

7. Conclusion

This thesis has demonstrated that Dutch TTO pupils started out with a larger vocabulary size in their L2 English and maintained this lead over time. Yet, while all pupils improved significantly in their vocabulary knowledge over time, with every school year outperforming the lower years, TTO pupils did not improve to a significantly higher degree compared to non-TTO pupils. TTO and non-TTO pupils scored similarly low on phonemic discrimination ability and did not improve over time, surprisingly enough.

The main research question of the present thesis was the following: What is the effect of bilingual secondary education on Dutch high school pupils' lexical decision and phonemic discrimination ability in English? In this study, TTO pupils outperformed non-TTO pupils in every school year on vocabulary scores, as was expected beforehand. However, it is not entirely certain that this is the effect of TTO rather than another factor entirely, such as TTO pupils' motivation to learn the target language or a high language aptitude, as the results provided no proof that TTO pupils acquire *more* vocabulary knowledge in comparison to pupils following the non-TTO track. Rather, TTO pupils seem to start out with a larger vocabulary size in high school. If TTO pupils had indeed acquired relatively more language skills than non-TTO pupils, there should have been an interaction effect between school year and TTO in our data, which was not found in any analysis conducted in this study. TTO pupils maintained their existing lead in vocabulary size over time, but did not seem to improve more strongly than non-TTO pupils. Additionally, TTO pupils did not score better at all on phonemic discrimination ability, regardless of what phonemic contrast (plosives, fricatives, vowels) was tested. Besides, no improvement on phonemic discrimination ability over time was found for either the TTO pupils or the non-TTO pupils, indicating that neither TTO nor standard language education affected this particular language skill at all. Surprisingly enough, the findings in this study imply that language education has no effect on pupils' ability to discriminate phonemes in their target language. Schools may need to pay more attention to pupils' acquisition of phonemic contrasts, for instance through explicit types of instruction, in order to draw the pupils' attention to the subtle phonemic differences. Possibly, such an approach may entail an improvement over time and pupils with TTO might benefit from the additional exposure in English they receive. The quality of this exposure may also play a role in pupils' ability to discriminate phonemic contrasts. It should be noted that in the current study is no way to be sure whether the pupils heard the phonemic differences and accepted them because they were used to phonemic variation, or if they were unable to discriminate. This is a limitation of this current thesis.

Additionally, the comparison between d-prime and accuracy revealed little difference in results for the most part, but a greater ability of the d-prime scores to explain statistical variance in most analyses. The d-prime scores have taken participants' response bias in consideration, and this inclusion should lead to more accurate results. Therefore, I believe that the d-prime should be the preferred dependent variable in further research.

This current thesis has aimed to contribute to a gap in literature about the possible benefits of bilingual education on phonemic discrimination and vocabulary size. In order to do so, we have studied TTO and non-TTO pupils in different school years and analysed their scores on a lexical decision task that measures both phonemic discrimination ability and vocabulary size. To summarise, this study did not reveal a particular advantage of TTO in the Netherlands; for that, perhaps, phonemic discrimination is too difficult to improve through exposure alone or there has not yet been enough (qualitatively sufficient) exposure for TTO and non-TTO pupils to improve this skill. TTO pupils outperformed non-TTO pupils on vocabulary size, but did not, in fact, demonstrate greater improvement in comparison to non-TTO pupils over time, which suggests that it is not TTO that causes this type of improvement. The question remains, in this case, if TTO is truly as beneficial as is often claimed. Researchers have wondered before whether children following bilingual education do not simply outperform others because of higher motivation (De Smet et al., 2019; Mearns, 2016) or because they start out with superior language skills (Bruton, 2011) and not necessarily because of the additional gains provided by TTO. The findings of this thesis lead me to conclude that, in order to establish the role that TTO has on pupils' language skills over time, more research is necessary. Especially phonological skills are still under-researched. Until then, it is important that we remain aware of the role of TTO on language skills and do not assume that all skills automatically improve because of any additional language exposure that TTO offers.

Bibliography

Abrahamsson, N., & Hyltenstam, K. (2009). Age of Onset and Nativelikeness in a Second Language: Listener Perception Versus Linguistic Scrutiny. *Language Learning*, *59*(2), 249–306. https://doi.org/10.1111/j.1467-9922.2009.00507.x

Abu-Rabia, S., & Kehat, S. (2004). The Critical Period for Second Language Pronunciation: Is there such a thing? Ten case studies of late starters who attained a native-like Hebrew accent SalimAbu-RabiaFaculty of EducationUniversity of HaifaMt. *Educational Psychology*.

Admiraal, W., Westhoff, G. J., & De Bot, K. (2006). Evaluation of bilingual secondary education in the Netherlands: Students' language proficiency in English. *Educational Research and Evaluation*, *12*. https://doi.org/10.1080/13803610500392160

Agustín Llach, M. P. (2017). The effects of the CLIL approach in young foreign language learners' lexical profiles. *International Journal of Bilingual Education and Bilingualism*, *20*(5), 557–573. https://doi.org/10.1080/13670050.2015.1103208

Alonso, E., Grisaleña, J., & Campo, A. (2008). Plurilingual education in secondary schools: Analysis of results. *International CLIL Research Journal*, *1*, 36–49.

Baker, W., & Trofimovich, P. (2005). Interaction of Native- and Second-Language Vowel System(s) in Early and Late Bilinguals. *Language and Speech*, *48*(1), 1–27. https://doi.org/10.1177/00238309050480010101

Bosch, L., Costa, A., & Sebastian Galles, N. (2000). First and second language vowel perception in early bilinguals. *European Journal of Cognitive Psychology*, *12*, 189–221. https://doi.org/10.1080/09541446.2000.10590222

Broersma, M. (2005). *Phonetic and lexical processing in a second language*. PhD Dissertation. https://repository.ubn.ru.nl/handle/2066/56388

Broersma, M. (2012). Increased lexical activation and reduced competition in second-language listening. *Language and Cognitive Processes*, *27*(7–8), 1205–1224. https://doi.org/10.1080/01690965.2012.660170

Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System*, *36*(1), 22–34. https://doi.org/10.1016/j.system.2007.11.003

Bruton, A. (2011). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System*, *39*(4), 523–532. https://doi.org/10.1016/j.system.2011.08.002

Bulon, A., & Meunier, F. (2020). Comparing CLIL and non-CLIL learners' phrasicon in L2 Dutch: The (expected) winner does not take it all. *International Journal of Bilingual Education and Bilingualism*, *0*(0), 1–24. https://doi.org/10.1080/13670050.2020.1834502

Campbell, J. M., & Dommestrup, A. K. (2010). Peabody Picture Vocabulary Test. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini Encyclopedia of Psychology* (1st ed., pp. 1–1). Wiley. https://doi.org/10.1002/9780470479216.corpsy0649

Daidone, D., & Darcy, I. (2021). Vocabulary Size Is a Key Factor in Predicting Second Language Lexical Encoding Accuracy. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.688356

Darcy, I., Park, H., & Yang, C.-L. (2015). Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences*, *40*, 63–72. https://doi.org/10.1016/j.lindif.2015.04.005

de Graaff, R. (1997). THE *EXPERANTO* EXPERIMENT: Effects of Explicit Instruction on Second Language Acquisition. *Studies in Second Language Acquisition*, *19*(2), 249–276. https://doi.org/10.1017/S0272263197002064

De Graaff, R. (2013). Graaff, R. de (2013). Taal om te leren: Iedere docent een CLIL-docent. Levende Talen Magazine, 100(7), 1-13. *Levende Talen Magazine*, *100*, 1–13.

De Schryver, J., Neijt, A., Ghesquière, P., & Ernestus, M. (2013). Zij surfde, maar hij durfte niet: De spellingproblematiek van de zwakke verleden tijd in Nederland en Vlaanderen. *Dutch Journal of Applied Linguistics*, *2*(2), 133–151. https://doi.org/10.1075/dujal.2.2.01de

De Smet, A., Mettewie, L., Hiligsmann, P., Galand, B., & Van Mensel, L. (2019). Does CLIL shape language attitudes and motivation? Interactions with target languages and instruction levels. *International Journal of Bilingual Education and Bilingualism*, *0*(0), 1–20. https://doi.org/10.1080/13670050.2019.1671308

DeKeyser, R. (2003). Implicit and Explicit Learning. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 312–348). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470756492.ch11

Díaz, B., Mitterer, H., Broersma, M., & Sebastián-Gallés, N. (2012). Individual differences in late bilinguals' L2 phonological processes: From acoustic-phonetic analysis to lexical access. *Learning and Individual Differences*, *22*(6), 680–689. https://doi.org/10.1016/j.lindif.2012.05.005

European Commission. (2006). *Content and Language Integrated Learning (CLIL) at school in Europe*. Eurydice.

Georgiou, G. P., Perfilieva, N. V., & Tenizi, M. (2020). Vocabulary Size Leads to Better Attunement to L2 Phonetic Differences: Clues from Russian Learners of English. *Language Learning and Development*, *16*(4), 382–398. https://doi.org/10.1080/15475441.2020.1814779

Gerritsen, M., Van Meurs, F., Planken, B., & Korzilius, H. (2016). A reconsideration of the status of English in the Netherlands within the Kachruvian Three Circles model. *World Englishes*, *35*(3), 457–474. https://doi.org/10.1111/weng.12206

Haatveit, B. C., Sundet, K., Hugdahl, K., Ueland, T., Melle, I., & Andreassen, O. A. (2010). The validity of d prime as a working memory index: Results from the “Bergen n-back task”. *Journal of Clinical and Experimental Neuropsychology*, *32*(8), 871–880. https://doi.org/10.1080/13803391003596421

Harley, B., & Hart, D. (1997). Language Aptitude and Second Language Proficiency in Classroom Learners of Different Starting Ages. *Studies in Second Language Acquisition*, *19*(3), 379–400.

Hayes-Harb, R., & Masuda, K. (2008). Development of the ability to lexically encode novel second language phonemic contrasts. *Second Language Research*, *24*(1), 5–33.

Hendrikx, I. (2019). *The acquisition of intensifying constructions in Dutch and English by French-speaking CLIL and non-CLIL students: Cross-linguistic influence and exposure effects* [UCL - Université Catholique de Louvain]. https://dial.uclouvain.be/pr/boreal/en/object/boreal%3A212617

Hommel, M. (2018). The Role of Orthography and Phoneme Inventory in Dutch Students’ Speech Perception in the EFL Classroom. *Philologia*, *16*(16), 65–75. https://doi.org/10.18485/philologia.2018.16.16.4

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, *19*(3), 227–245. https://doi.org/10.1191/0265532202lt229oa

Huijbregtse, I. (2001). Onderwijs in twee talen. *Levende Talen Tijdschrift*, *2*(1), Article 1.

Hulstijn, J. H. (2012). Incidental Learning in Second Language Acquisition. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (p. wbeal0530). Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0530

Jiménez Catalán, R. M., & Agustín Llach, M. P. (2017). CLIL or time? Lexical profiles of CLIL and non-CLIL EFL learners. *System*, *66*, 87–99. https://doi.org/10.1016/j.system.2017.03.016

Kissling, E. M. (2014). What Predicts the Effectiveness of Foreign-Language Pronunciation Instruction? Investigating the Role of Perception and Other Individual Differences. *The Canadian Modern Language Review*, *70*(4), 532–558. https://doi.org/10.3138/cmlr.2161

Kissling, E. M. (2015). Phonetics instruction improves learners’ perception of L2 sounds. *Language Teaching Research*, *19*(3), 254–275. https://doi.org/10.1177/1362168814541735

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), Article 11. https://doi.org/10.1038/nrn1533

Lacabex, E. G., & Gallardo-del-Puerto, F. (2020). Explicit phonetic instruction vs. implicit attention to native exposure: Phonological awareness of English schwa in CLIL. *International Review of Applied Linguistics in Language Teaching*, *58*(4), 419–442. https://doi.org/10.1515/iral-2017-0079

Lasagabaster, D., & Doiz, A. (2017). A Longitudinal Study on the Impact of CLIL on Affective Factors. *Applied Linguistics*, *38*, amv059. https://doi.org/10.1093/applin/amv059

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Lersveen, L.-R. (2018). The perception and production of nonnative English consonants in native Norwegian speakers [Master thesis, NTNU]. In *49*. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2574550

Llompart, M. (2019). *Bridging the gap between phonetic abilities and the lexicon in second language learning* [Text.PhDThesis, Ludwig-Maximilians-Universität München]. https://edoc.ub.uni-muenchen.de/24192/

Llompart, M. (2021a). Lexical and Phonetic Influences on the Phonolexical Encoding of Difficult Second-Language Contrasts: Insights From Nonword Rejection. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.659852

Llompart, M. (2021b). Phonetic categorization ability and vocabulary size contribute to the encoding of difficult second-language phonological contrasts into the lexicon. *Bilingualism: Language and Cognition*, *24*(3), 481–496. https://doi.org/10.1017/S1366728920000656

Llompart, M., & Reinisch, E. (2017). Articulatory information helps encode lexical contrasts in a second language. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 1040–1056. https://doi.org/10.1037/xhp0000383

Llompart, M., & Reinisch, E. (2018). Robustness of phonolexical representations relates to phonetic flexibility for difficult second language sound contrasts. *Bilingualism: Language and Cognition*, *22*, 1085–1100. https://doi.org/10.1017/S1366728918000925

Llompart, M., & Reinisch, E. (2019). Robustness of phonolexical representations relates to phonetic flexibility for difficult second language sound contrasts. *Bilingualism: Language and Cognition*, *22*(5), 1085–1100. https://doi.org/10.1017/S1366728918000925

Lorenzo, F., Casal, S., & Moore, P. (2010). The Effects of Content and Language Integrated Learning in European Education: Key Findings from the Andalusian Bilingual Sections Evaluation Project. *Applied Linguistics*, *31*(3), 418–442. https://doi.org/10.1093/applin/amp041

Mearns, T. (2016). *Chicken, Egg or a Bit of Both? Motivation in bilingual education (TTO) in the Netherlands.* https://doi.org/10.13140/RG.2.1.4205.9926

Olsson, E. (2021). A comparative study of CLIL implementation in upper secondary school in Sweden and students' development of L2 English academic vocabulary. *Language Teaching Research*, 13621688211045000. https://doi.org/10.1177/13621688211045000

Paradis, J. (2011). Individual differences in child English second language acquisition: Comparing child-internal and child-external factors. *Linguistic Approaches to Bilingualism*, *1*(3), 213–237. https://doi.org/10.1075/lab.1.3.01par

Pérez Cañado, M. (2018). CLIL and Educational Level: A Longitudinal Study on the Impact of CLIL on Language Outcomes. *Porta Linguarum*, *2018*. https://doi.org/10.30827/Digibug.54022

Pinget, A.-F., Kager, R., & Van de Velde, H. (2020). Linking Variation in Perception and Production in Sound Change: Evidence from Dutch Obstruent Devoicing. *Language and Speech*, *63*(3), 660–685. https://doi.org/10.1177/0023830919880206

Ruiz de Zarobe, Y., & Lasagabaster, D. (2010). CLIL in a Bilingual Community: The Basque Autonomous Community. *CLIL in Spain: Implementation, Results and Teacher Training*, 12–29.

Rumlich, D. (2018). Current research on CLIL and bilingual education in the Netherlands: A discussion. *Dutch Journal of Applied Linguistics*, *7*(2), 264–273. https://doi.org/10.1075/dujal.00003.rum

Saito, K. (2015). The Role Of Age Of Acquisition In Late Second Language Oral Proficiency Attainment. *Studies in Second Language Acquisition*, *37*(4), 713–743. https://doi.org/10.1017/S0272263115000248

Schepens, J. J., Hout, R. W. N. M. van, & Slik, F. W. P. van der. (2022). Linguistic dissimilarity increases age-related decline in adult language learning. *Studies in Second Language Acquisition*, 1–22. https://doi.org/10.1017/S0272263122000067

Seikkula-Leino, J. (2007). CLIL Learning: Achievement Levels and Affective Factors. *Language and Education*, *21*, 328–341. https://doi.org/10.2167/le635.0

Thomson, R. I. (2018). High Variability [Pronunciation] Training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation*, *4*(2), 208–231. https://doi.org/10.1075/jslp.17038.tho

Thorin, J., Sadakata, M., Desain, P., & McQueen, J. M. (2018). Perception and production in interaction during non-native speech category learninga). *The Journal of the Acoustical Society of America*, *144*(1), 92. https://doi.org/10.1121/1.5044415

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2022). The Effects of Talker Variability and Frequecy of Exposure on the Acquisition of Spoken Word Knowledge. *Studies in Second Language Acquisition*, *44*(2), 357–380. https://doi.org/10.1017/S0272263121000218

Van de Velde, H., & Van Hout, R. (2001). The devoicing of fricatives in a reading task. *Linguistics in the Netherlands*, *18*, 219–229. https://doi.org/10.1075/avt.18.22van

Van Kampen, E., Admiraal, W., & Berry, A. (2018). Content and language integrated learning in the Netherlands: Teachers' self-reported pedagogical practices. *International Journal of Bilingual Education and Bilingualism*, *21*(2), 222–236. https://doi.org/10.1080/13670050.2016.1154004

Van Leussen, J.-W., & Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology*, *6*. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01000

Van ’t Erve, D. (2021, December 7). *Onderwijsraad wil publiek onderwijs beter beschermen*. De Algemene Onderwijsbond. https://www.aob.nl/nieuws/scholen-onvoldoende-bewust-van-risicos-privatisering/

Verspoor, M., De Bot, K., & Xu, X. (2015). The effects of English bilingual education in the Netherlands. *Journal of Immersion and Content-Based Language Education*, *3*, 4–27. https://doi.org/10.1075/jicb.3.1.01ver

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*(1), 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0

Wolfswinkler, K., & Reinisch, E. (2016). The impact of accent familiarity on the perception of difficult sound contrasts for German learners of English. *Tagung Phonetik Und Phonologie Im Deutschsprachigen Raum*, *12*, 232–236.

Ziegler, J. C., Muneaux, M., & Grainger, J. (2003). Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language*, *48*(4), 779–793. https://doi.org/10.1016/S0749-596X(03)00006-8

Appendices

Appendix 1. Word list for phonemic discrimination ability

list	condition	item	CV	Sound	Item number	Trial Number
1	888	vote	999	999	888	888
1	888	make	999	999	888	888
1	888	trink	999	999	888	888
1	888	dull	999	999	888	888
1	888	blonce	999	999	888	888
1	888	vaidge	999	999	888	888
1	888	plog	999	999	888	888
1	888	town	999	999	888	888
1	888	drice	999	999	888	888
1	888	spend	999	999	888	888
1	3	shade	999	999	79	1
1	1	shave	1	2	3	2
1	3	dish	999	999	94	3
1	4	shorp	999	999	186	4
1	4	yearl	999	999	141	5
1	1	fresh	2	9	20	6
1	3	chair	999	999	98	7
1	3	steam	999	999	80	8
1	3	crane	999	999	127	9
1	1	laugh	1	5	10	10
1	2	groof	1	2	36	11
1	3	glaze	999	999	93	12
1	4	breen	999	999	168	13
1	1	praise	1	3	5	14
1	4	fub	999	999	138	15
1	3	judge	999	999	72	16
1	2	bleck	2	10	58	17

list	condition	item	CV	Sound	Item number	Trial Number
1	1	trap	2	10	26	18
1	4	shipe	999	999	137	19
1	3	spouse	999	999	118	20
1	3	soak	999	999	90	21
1	1	cube	1	1	1	22
1	4	wath	999	999	157	23
1	3	crumb	999	999	105	24
1	3	shine	999	999	131	25
1	1	choice	1	7	14	26
1	3	grape	999	999	87	27
1	3	game	999	999	76	28
1	4	noik	999	999	148	29
1	1	sharp	1	8	16	30
1	4	shoul	999	999	158	31
1	3	ship	999	999	107	32
1	1	slam	2	10	29	34
1	3	boil	999	999	128	35
1	2	swat	2	9	54	36
1	4	stroil	999	999	192	37
1	4	neve	999	999	182	38
1	1	plank	2	10	31	39
1	3	south	999	999	102	40
1	3	vase	999	999	114	41
1	4	clil	999	999	151	42
1	2	glope	1	1	33	43
1	4	hoke	999	999	177	44
1	4	crale	999	999	133	45
1	3	street	999	999	122	46
1	4	dawsh	999	999	147	47
1	4	hean	999	999	200	48

list	condition	item	CV	Sound	Item number	Trial Number
1	4	fraf	999	999	174	49
1	1	beard	1	4	7	50
1	3	dread	999	999	123	51
1	4	parve	999	999	152	52
1	3	growl	999	999	95	53
1	1	smash	2	10	28	54
1	4	firp	999	999	161	55
1	3	chap	999	999	126	56
1	4	chadge	999	999	145	57
1	1	chess	2	9	24	58
1	4	stirl	999	999	135	59
1	4	feuth	999	999	143	60
1	4	rouch	999	999	166	61
1	1	skirt	1	6	11	62
1	2	walth	2	9	53	63
1	4	deadge	999	999	149	64
1	3	crate	999	999	110	65
1	1	bank	2	10	25	66
1	3	farm	999	999	82	67
1	2	moof	1	2	35	68
1	3	youth	999	999	69	69
1	4	pute	999	999	144	70
1	1	spank	2	10	32	71
1	3	scare	999	999	111	72
1	3	search	999	999	100	73
1	4	sprull	999	999	189	74
1	4	strawn	999	999	164	75
1	1	bench	2	9	22	76
1	2	dask	2	9	51	77
1	4	brile	999	999	197	79

list	condition	item	CV	Sound	Item number	Trial Number
1	3	drowse	999	999	75	80
1	4	parf	999	999	183	81
1	3	flush	999	999	66	82
1	3	stir	999	999	129	83
1	3	touch	999	999	68	84
1	1	news	1	3	6	85
1	4	kiff	999	999	154	86
1	3	mouse	999	999	67	87
1	4	sporf	999	999	198	88
1	4	spudge	999	999	190	89
1	2	cheece	1	3	37	90
1	3	bath	999	999	96	91
1	4	lir	999	999	180	92
1	3	broom	999	999	109	93
1	2	blenk	2	10	61	94
1	1	voice	1	7	13	95
1	4	swut	999	999	163	96
1	4	fluss	999	999	173	97
1	2	stiv	1	5	41	98
1	3	drake	999	999	85	99
1	4	jark	999	999	179	100
1	4	lon	999	999	136	101
1	2	shib	1	8	48	102
1	3	foam	999	999	124	103
1	3	shrill	999	999	103	104
1	1	cheap	1	8	15	105
1	2	sheeb	1	8	47	106
1	3	crown	999	999	92	107
1	3	nice	999	999	101	108
1	4	squayle	999	999	156	109

list	condition	item	CV	Sound	Item number	Trial Number
1	2	dwarve	1	5	42	110
1	1	dive	1	2	4	111
1	4	mome	999	999	181	112
1	4	spetch	999	999	187	113
1	2	quast	2	9	56	114
1	4	dreeve	999	999	171	115
1	3	gleam	999	999	121	116
1	4	quirze	999	999	185	117
1	2	jat	2	9	55	118
1	4	gleathe	999	999	188	119
1	3	full	999	999	89	120
1	3	suck	999	999	73	121
1	2	prout	1	4	39	122
1	3	mole	999	999	88	124
1	3	guide	999	999	97	125
1	2	phrace	1	3	38	126
1	1	scarf	1	5	9	127
1	3	sneeze	999	999	116	128
1	4	trorse	999	999	195	129
1	2	scretch	2	10	60	130
1	3	cruel	999	999	120	131
1	4	hidge	999	999	142	132
1	4	drile	999	999	175	133
1	1	dress	2	9	18	134
1	4	strit	999	999	193	135
1	4	drune	999	999	159	136
1	3	skill	999	999	84	137
1	2	kiz	1	7	45	138
1	4	vabe	999	999	140	139
1	3	jeer	999	999	65	140

list	condition	item	CV	Sound	Item number	Trial Number
1	3	meek	999	999	83	141
1	2	grem	2	10	63	142
1	4	trave	999	999	194	143
1	3	dare	999	999	125	144
1	4	boin	999	999	134	145
1	2	flied	1	6	43	146
1	1	blade	1	4	8	147
1	4	meach	999	999	165	148
1	4	grish	999	999	178	149
1	3	snake	999	999	77	150
1	1	press	2	9	17	151
1	3	frown	999	999	91	152
1	3	thin	999	999	99	153
1	2	splesh	2	10	62	154
1	1	breast	2	9	19	155
1	4	vike	999	999	155	156
1	4	brear	999	999	167	157
1	2	smal	2	9	50	158
1	1	swell	2	9	23	159
1	4	hube	999	999	153	160
1	4	shrit	999	999	139	161
1	2	thenk	2	10	57	162
1	1	smart	1	6	12	163
1	4	chert	999	999	170	164
1	3	screen	999	999	86	165
1	2	tupe	1	1	34	166
1	1	chest	2	9	21	167
1	4	vodge	999	999	196	169
1	2	prem	2	10	64	170
1	3	lash	999	999	112	171

list	condition	item	CV	Sound	Item number	Trial Number
1	4	brong	999	999	169	172
1	4	clum	999	999	172	173
1	2	renk	2	10	59	174
1	3	rage	999	999	106	175
1	3	church	999	999	117	176
1	3	sake	999	999	81	177
1	2	nurze	1	7	46	178
1	4	mosh	999	999	191	179
1	3	duck	999	999	78	180
1	3	pipe	999	999	70	181
1	2	glite	1	4	40	182
1	4	froop	999	999	176	183
1	3	chief	999	999	71	184
1	4	waph	999	999	160	185
1	2	dath	2	9	49	186
1	4	sman	999	999	162	187
1	1	rub	1	1	2	188
1	4	plorn	999	999	184	189
1	2	spid	1	6	44	190
1	3	yearn	999	999	115	191
1	3	weep	999	999	132	192
1	3	fight	999	999	130	193
1	4	murp	999	999	199	194
1	1	lamp	2	10	27	195
1	2	brath	2	9	52	196
1	3	blaze	999	999	113	197
1	3	scheme	999	999	108	198
1	4	trif	999	999	150	199
1	3	crook	999	999	119	200

Condition is the word category: 1 = real word, 2 = near word, 3 = filler word, 4 = filler non-word.

CV states whether the manipulated item (whether present or not) is a consonant (=1) or vowel (=2)

Sound defines the specific sound that is being manipulated.

Item number refers to the position of the item when they were originally paired.

Trial number is the item number in the test.

Appendix 2. Word list for lexical decision ability

list	condition	item	CV	Sound	Item number	Trial Number
2	888	vote	999	999	888	888
2	888	make	999	999	888	888
2	888	trink	999	999	888	888
2	888	dull	999	999	888	888
2	888	blonce	999	999	888	888
2	888	vaidge	999	999	888	888
2	888	plog	999	999	888	888
2	888	town	999	999	888	888
2	888	drice	999	999	888	888
2	888	spend	999	999	888	888
2	4	boin	999	999	134	1
2	4	vabe	999	999	140	2
2	3	meek	999	999	83	3
2	4	dawsh	999	999	147	4
2	2	slem	2	10	61	5
2	3	ship	999	999	107	6
2	4	shoul	999	999	158	7
2	1	blank	2	10	29	8
2	2	drass	2	9	50	9
2	4	lir	999	999	180	10
2	3	mouse	999	999	67	11
2	4	spetch	999	999	187	12
2	2	trep	2	10	58	13
2	4	vike	999	999	155	14
2	3	steam	999	999	80	15
2	3	dread	999	999	123	16
2	4	yearl	999	999	141	18
2	3	sake	999	999	81	19
2	3	growl	999	999	95	20

list	condition	item	CV	Sound	Item number	Trial Number
2	2	scarve	1	5	41	21
2	1	scratch	2	10	28	22
2	4	clum	999	999	172	23
2	3	spouse	999	999	118	24
2	2	smard	1	6	44	25
2	1	thank	2	10	25	26
2	3	lash	999	999	112	27
2	3	blood	999	999	74	28
2	2	newce	1	3	38	29
2	1	smell	2	9	18	30
2	4	shorp	999	999	186	31
2	3	chap	999	999	126	32
2	2	sharb	1	8	48	33
2	1	splash	2	10	30	34
2	4	mosh	999	999	191	35
2	3	chief	999	999	71	36
2	3	crane	999	999	127	37
2	3	thin	999	999	99	38
2	3	touch	999	999	68	39
2	4	rouch	999	999	166	40
2	3	fine	999	999	124	41
2	4	swut	999	999	163	42
2	4	firp	999	999	161	43
2	3	scheme	999	999	108	44
2	3	rage	999	999	106	46
2	4	spudge	999	999	190	47
2	4	plorn	999	999	184	48
2	3	church	999	999	117	49
2	4	fraf	999	999	174	50
2	1	death	2	9	17	51

list	condition	item	CV	Sound	Item number	Trial Number
2	2	skird	1	6	43	52
2	3	fight	999	999	130	53
2	4	hoke	999	999	177	54
2	4	pute	999	999	144	55
2	2	rup	1	1	34	56
2	1	rank	2	10	27	57
2	3	full	999	999	89	58
2	4	jark	999	999	179	59
2	4	froop	999	999	176	60
2	1	sweat	2	9	22	61
2	2	cupe	1	1	33	62
2	4	drile	999	999	175	63
2	3	sneeze	999	999	116	64
2	1	breath	2	9	20	65
2	4	shipe	999	999	137	66
2	3	scare	999	999	111	67
2	4	parve	999	999	152	68
2	1	groove	1	2	4	69
2	4	sman	999	999	162	70
2	2	prass	2	9	49	71
2	3	drowse	999	999	75	72
2	3	bath	999	999	96	74
2	3	south	999	999	102	75
2	4	waph	999	999	160	76
2	4	hidge	999	999	142	77
2	4	chadge	999	999	145	78
2	1	glide	1	4	8	79
2	3	snake	999	999	77	80
2	2	praice	1	3	37	81
2	3	frown	999	999	91	82

list	condition	item	CV	Sound	Item number	Trial Number
2	4	noik	999	999	148	83
2	4	vodge	999	999	196	84
2	2	lemp	2	10	59	85
2	4	strawn	999	999	164	86
2	1	pram	2	10	32	87
2	3	drake	999	999	85	88
2	3	shade	999	999	79	89
2	3	grape	999	999	87	90
2	3	crumb	999	999	105	91
2	2	chass	2	9	56	92
2	3	farm	999	999	82	93
2	1	gram	2	10	31	94
2	3	shrill	999	999	103	95
2	3	stir	999	999	129	96
2	2	choise	1	7	46	97
2	4	lon	999	999	136	98
2	1	jet	2	9	23	99
2	4	stroil	999	999	192	100
2	3	screen	999	999	86	102
2	1	flight	1	6	11	103
2	3	nice	999	999	101	104
2	4	mome	999	999	181	105
2	2	cheab	1	8	47	106
2	1	tube	1	1	2	107
2	3	street	999	999	122	108
2	3	skill	999	999	84	109
2	2	beart	1	4	39	110
2	1	cheese	1	3	5	111
2	4	neve	999	999	182	112
2	3	crown	999	999	92	113

list	condition	item	CV	Sound	Item number	Trial Number
2	4	wath	999	999	157	114
2	1	dwarf	1	5	10	115
2	3	game	999	999	76	116
2	4	chert	999	999	170	117
2	3	youth	999	999	69	118
2	2	voise	1	7	45	119
2	1	desk	2	9	19	120
2	3	crate	999	999	110	121
2	4	sprull	999	999	189	122
2	3	cruel	999	999	120	123
2	1	spit	1	6	12	124
2	4	sporf	999	999	198	125
2	4	fluss	999	999	173	126
2	2	banch	2	9	54	127
2	1	globe	1	1	1	128
2	4	strit	999	999	193	130
2	3	glaze	999	999	93	131
2	1	move	1	2	3	132
2	4	squayle	999	999	156	133
2	4	deadge	999	999	149	134
2	2	frash	2	9	52	135
2	4	dreeve	999	999	171	136
2	4	brile	999	999	197	137
2	3	shine	999	999	131	138
2	1	proud	1	4	7	139
2	4	hean	999	999	200	140
2	2	plenk	2	10	63	141
2	4	grish	999	999	178	142
2	1	ship	1	8	16	143
2	4	stirl	999	999	135	144

list	condition	item	CV	Sound	Item number	Trial Number
2	2	spen	2	10	62	145
2	4	breen	999	999	168	146
2	4	quirze	999	999	185	147
2	1	nurse	1	7	14	148
2	2	swall	2	9	55	149
2	4	shrit	999	999	139	150
2	3	pipe	999	999	70	151
2	3	dish	999	999	94	152
2	1	wealth	2	9	21	153
2	3	curl	999	999	104	154
2	2	shafe	1	2	35	155
2	3	mole	999	999	88	156
2	3	duck	999	999	78	158
2	2	dife	1	2	36	159
2	4	brong	999	999	169	160
2	4	murp	999	999	199	161
2	3	yearn	999	999	115	162
2	2	lauve	1	5	42	163
2	3	vase	999	999	114	164
2	1	phrase	1	3	6	165
2	3	blaze	999	999	113	166
2	4	brear	999	999	167	167
2	4	trorse	999	999	195	168
2	4	feuth	999	999	143	169
2	3	broom	999	999	109	170
2	3	dare	999	999	125	171
2	1	black	2	10	26	172
2	2	brast	2	9	51	173
2	4	trave	999	999	194	174
2	4	hube	999	999	153	175

list	condition	item	CV	Sound	Item number	Trial Number
2	3	crook	999	999	119	176
2	3	gleam	999	999	121	177
2	2	blate	1	4	40	178
2	4	fub	999	999	138	179
2	4	crale	999	999	133	180
2	3	jeer	999	999	65	181
2	2	chast	2	9	53	182
2	4	clil	999	999	151	183
2	1	sheep	1	8	15	184
2	2	spenk	2	10	64	186
2	3	boil	999	999	128	187
2	3	flush	999	999	66	189
2	2	smesh	2	10	60	190
2	4	brutt	999	999	146	191
2	3	weep	999	999	132	192
2	4	parf	999	999	183	193
2	4	gleathe	999	999	188	194
2	1	quest	2	9	24	195
2	4	trif	999	999	150	196
2	2	benk	2	10	57	197
2	3	judge	999	999	72	198
2	3	soak	999	999	90	199
2	1	kiss	1	7	13	200

Condition is the word category: 1 = real word, 2= near word, 3 = filler word, 4 =filler non-word.

CV states whether the manipulated item (whether present or not) is a consonant (=1) or vowel (=2).

Sound defines the specific sound that is being manipulated.

Item number refers to the position of the item when they were originally paired.

Trial number is the item number in the test.