The use of native speaker norms in critical period hypothesis research

Andringa, S.

DOI
10.1017/S0272263113000600
Publication date
2014
Document Version
Final published version
Published in
Studies in Second Language Acquisition

Studies in Second Language Acquisition, 2014, **36**, 565–596.
doi:10.1017/S0272263113000600

THE USE OF NATIVE SPEAKER NORMS IN CRITICAL PERIOD HYPOTHESIS RESEARCH

Sible Andringa
University of Amsterdam

In critical period hypothesis (CPH) research, native speaker (NS) norm groups have often been used to determine whether nonnative speakers (NNSs) were able to score within the NS range of scores. One goal of this article is to investigate what NS samples were used in previous CPH research. The literature review shows that NS control groups tend to be small and highly educated and that detailed background information is usually not provided. Another goal of this article is to investigate how the NS norm group may affect the incidence of nativelike performance by NNSs. To this end, 124 NSs and 118 NNSs of Dutch completed five comprehension tasks and a vocabulary task. On the basis of mean scores and standard deviations, norms were determined for a representative and a nonrepresentative (highly educated) subsample of NSs. Also, separate norms were constructed for the high- and low-frequency items within a task. Exact McNemar tests were used to establish that the incidence of nativelike performance by NNSs was significantly higher if a representative sample norm was used. The results also showed that, insofar as there were effects of frequency, norms based on low-frequency test items tended to be more inclusive. The results imply that the selection of NSs in CPH research deserves more consideration than it has

This research was funded by the Netherlands Organisation for Scientific Research by a grant awarded to Jan Hulstijn (NWO grant 360–70-230).

I would like to express my sincere gratitude to Nomi Olsthoorn and Jan Hulstijn for their contributions in developing the materials presented in this study, to Tineke van der Linde and Netta Meijer for collecting and organizing the data, and to Dirk Vet for his technical assistance. I would also like to thank Catherine van Beuningen, Jan Hulstijn, Rob Schoonen, and four anonymous *SSLA* reviewers for their constructive feedback on this text.

Correspondence concerning this article should be addressed to Sible Andringa, Center for Language and Communication, University of Amsterdam, Spuistraat 134, 1012 VB, Amsterdam, the Netherlands. E-mail: S.J.Andringa@uva.nl

received in the past; they also suggest that NS ceiling performance is potentially useful in determining nativelike performance.

One of the most fundamental issues in the field of SLA concerns the (non)existence of a critical period for language learning: Can late second language (L2) learners ever achieve nativelike levels of mastery in the L2? The critical period hypothesis (CPH) for language learning was proposed in 1959 by Penfield and Roberts, and later by Lenneberg (1967). In its strongest form, the CPH predicts that late L2 learners cannot achieve a nativelike level of proficiency because language learning is fundamentally different after a certain age (e.g., Bley-Vroman, 1989; Gregg, 1996; Long & Robinson, 1998). Weaker versions of the CPH acknowledge that L2 learning is affected by age effects, making nativelike attainment less likely for later onsets of learning, but nativelikeness is sometimes possible given the right circumstances (Birdsong & Molis, 2001; Bongaerts, Van Summeren, Planken, & Schils, 1997). Other versions of the CPH hold that nativelike levels of attainment may be possible but state that the learning mechanisms underlying early and late L2 acquisition are not the same. These versions of the CPH contend that nativelike levels can only be attained through more explicit types of learning because the ability to learn implicitly is gradually lost (DeKeyser, 2000). By now, many studies have been published that address aspects of the CPH (for reviews see DeKeyser, 2012; Hyltenstam & Abrahamsson, 2003), and these studies often include a native speaker (NS) sample. The goal of this study is to evaluate how NSs have been used in CPH research and to assess how the choice of NSs may have affected the results of such studies. First, I review if and how NSs have been used to falsify claims of the CPH, who the NSs in CPH research were, and how variable NS performance may be. Then, I present data demonstrating how the selection of the NS sample may affect the outcomes of CPH research.

One approach to falsification of the CPH does not involve NSs: Many have considered the correlation between age of onset (AoO) and L2 proficiency measures (e.g., Abrahamsson & Hyltenstam, 2009; Bialystok & Miller, 1999; DeKeyser, 2000). A strong correlation (negative or positive, depending on whether accuracy or speed measures are used) means that early onset is related to higher levels of ultimate attainment. However, DeKeyser (2012) pointed out that such correlations are not in support of the CPH unless they show the noncontinuity of the relationship between AoO and proficiency: Before the critical period, the relationship between AoO and achievement should be strong; after the critical period, the effect should cease to exist. DeKeyser's review showed that, in the vast majority of studies that provide relevant information, the effect of AoO does level off somewhere in puberty. Although this finding

is in support of the CPH because it suggests a qualitative change in the ability to learn the L2, it merely proves that L2 acquisition is subject to age effects (DeKeyser, 2012).

Another approach to falsification of the CPH is to demonstrate that the proficiency of a particular L2 learner is indistinguishable from that of NSs. In fact, Long (1990) suggested that finding one such learner would suffice to falsify the strong version of the CPH, which predicts that nativelike levels cannot be achieved. Abrahamsson and Hyltenstam (2009) reviewed most of the extant CPH work and found considerable variation in the reported incidence rates of nativelikeness, from none (e.g., Johnson & Newport, 1989) to rates toward 75% of a study's participant sample (e.g., Birdsong, 1992). Several researchers have pointed to the fact that differences in design may have caused this variation in outcomes (e.g., Abrahamsson & Hyltenstam, 2009; DeKeyser, 2013; Long, 2005). The incidence of nativelikeness obtained in any study is the result of (a) which L2 learners are selected, (b) which target structures and test tasks are used, and (c) which NSs are selected. Not only do the choices that are made in this regard determine the yardstick by which L2 learners' performance is gauged, but they actually reflect how the construct of nativelike mastery has been operationalized. The selection of L2 learners and study design choices has been subject to discussion; the selection of the NS controls has received little attention in the literature.

When it comes to the first aspect—the selection of the L2 sample—the majority of researchers seem to agree that it is best to select highly advanced learners (e.g., Hyltenstam & Abrahamsson, 2003; White & Genesee, 1996). Abrahamsson and Hyltenstam (2009), who presented probably the most sophisticated CPH study so far, tried to select only those L2 learners whose speech could not be perceived by NS judges as being nonnative. In their review of the literature, they illustrated that many studies explicitly stated that they selected advanced learners and provided information about length of residence, amount of exposure, and some other background variables. However, they also noted that many studies were unclear about how advanced their learners were in terms of L2 proficiency and that many samples may well have included L2 learners who were easily identifiable as nonnative. Long (2005) pointed to another crucial requirement with regard to the selection of appropriate L2 learners for CPH research: The L2 should have been acquired clearly past the hypothesized critical period. There still is considerable debate about where the end of the critical period lies, and some have claimed that there are multiple critical periods for different domains of language (Long, 1990). This is supported by the findings of Granena and Long (2013), who reported discontinuities in the correlation between AoO and performance, first for phonology, then for lexis and collocation, and finally for morphosyntax. In this light, it seems that

it is best to select learners well past their midteens, especially if they are tested on multiple aspects of proficiency. An additional concern was raised by Muñoz (2008), who points out that it may not be fair to apply the concept of nativelike attainment to L2 learners in instructed settings because the nature of the input received in these settings is such that nativelikeness cannot be expected. In other words, the selection of the L2 sample may also have to be motivated by the nature and the amount of the learners' experience with the L2.

When it comes to the choice of target structures and test tasks (the second aspect that determines a study's obtained incidence rate), a crucial question was raised by Birdsong (2005): "What is the range of linguistic behaviors that should be considered for falsification of the CPH/L2A [the critical period hypothesis for second language acquisition]?" (p. 322). DeKeyser's (2012) review shows that the existing body of research is largely limited to grammatical competence and pronunciation. Studies focusing on grammatical competence have often used grammaticality judgment tasks pertaining to one or several rules of grammar. Studies about pronunciation mostly involve holistic ratings of nativelikeness by NSs, although acoustic analysis techniques have also been used (e.g., B. Lee, Guion, & Harada, 2006). Hyltenstam and Abrahamsson (2003) stated that it is not valid to draw conclusions about critical periods on the basis of investigations into just one or two aspects of the L2, and they insisted that L2 learners should be subjected to in-depth linguistic scrutiny, which means that a multitask approach should be used so that many different aspects of L2 proficiency are investigated. This diminishes the possibility that the L2 learner happens to be nativelike in the domain of language that was tested but not in other domains. Additionally, Hyltenstam and Abrahamsson argued that, for statistical reasons, the tests used should not exhibit ceiling effects in NS performance. Differences between NSs and nonnative speakers (NNSs) may then be due to insufficient variation in NS performance. Abrahamsson and Hyltenstam (2009) adopted these recommendations and found that none of the late learners fell within the NS range, but this was also true for many of the early L2 learners.

Several researchers have claimed that criteria for testing the CPH should be derived from theories of language (Birdsong, 2005; Eubank & Gregg, 1999; Hyltenstam & Abrahamsson, 2003). In studies conducted within the Universal Grammar framework, grammatical structures were selected that required parameter (re)setting for learners of a particular first language (L1; e.g., Johnson & Newport, 1989). However, this limits the choice to only those features whose development is indeed subject to Universal Grammar. More recently, Hulstijn (2011) suggested a view on language proficiency that can potentially provide a theoretically motivated criterion for falsifying the CPH. He proposed the constructs of basic and higher language cognition. Basic language cognition refers

to those features of the language that are shared by all NSs of the language and that can be applied automatically. Anything beyond basic language cognition is higher language cognition and falls outside the realm of what it means to be nativelike. Although the suggestion is interesting, as things stand, the theory cannot be used to test the CPH, as it has not yet been verified empirically.

Decisions about what to test seem to be based mostly on empirical rather than theoretical grounds. Hyltenstam and Abrahamsson (2003) advocated a more empirical approach when they claimed that inquiry into the CPH should not be limited to core aspects of grammar. They pointed out that L2 learners should be tested on a wide array of language features to make sure that NNSs are nativelike in each and every domain of the L2, not just in one. Tests should include peripheral and language-specific features, such as knowledge of metaphors and idiomatic constructions, as these are the domains in which NSs and NNSs are likely to differ. Birdsong (2005) pointed out, however, that such an approach renders the CPH unfalsifiable. Even if L2 learners pass all tests, there is always the possibility that the one test that can identify them as NNSs is not part of the test battery. This leaves CPH researchers in a paradoxical situation: One can and should test L2 learners on as many aspects of language as possible, but it might never be enough to falsify the CPH.

The selection of the NS controls is the third aspect that determines whether NNSs can perform within the NS range. This aspect seems to have received very little consideration in the literature so far. To be more precise, few have raised concerns about the actual makeup of the NS norm groups, but the practice of comparing NSs and NNSs itself has not remained unchallenged within the CPH debate. The most important objection raised is that CPH research generally involves the comparison of monolingual and bilingual speakers (e.g., Birdsong, 2005; Cook, 2002; Grosjean, 1998). There is abundant evidence that bilingualism may affect both the L1 and the L2 systems (see Montrul, 2008) as well as more general cognitive processes (see Bialystok, 2005). Because the language systems of bilinguals are connected in intricate ways, it may be impossible by definition for L2 learners to have a L2 system that is identical to that of monolingual L1 speakers. As long as the L1 is active, it will exert some influence on the L2, even if the L2 is the dominant language. In a recent review, Hulstijn (2012) made specific recommendations concerning the selection of NSs for comparison with bilinguals. Most important, he recommended that NSs should not be more advanced in terms of the factors that shape linguistic experience, such as literacy and level of education.

Comparisons of NSs and NNSs have often been undertaken despite concerns about the validity of such comparisons. An inspection of the NS groups used in CPH studies shows that they are generally not in line with Hulstijn's recommendation. Table 1 lists 35 studies that tried to establish whether NNSs were able to score within the range of a group

of NS controls. It documents both the NS and NNS samples used in these studies; for the NSs, it also documents whatever was reported about the background of the NS sample used. Several things stand out. First, 2 studies did not include any NSs; they compared early and late learners only. The remaining 33 studies all made use of a NS group, although some referred to the norms provided by other studies (e.g., Birdsong & Molis, 2001; DeKeyser, 2000) instead of testing NSs themselves. In cases in which pronunciation was tested, the studies also included NS judges, which are usually small groups of expert language users (e.g., teachers). Second, the NS samples were often quite small: 14 studies made use of a sample of 10 speakers or fewer. Third, and perhaps most important, very little background information was provided in terms of NS literacy levels and levels of education: 21 out of 33 studies (64%) reported the group's mean age or whether they spoke with or without an accent, but further information about the NSs tested was not given. When information was provided, the NS group was invariably highly educated. It is fair to conclude that NSs were merely mentioned in passing in many studies. Finally, in some studies, NSs were selected to behave according to a certain standard, and speakers were removed from the sample because they did not meet this standard. For example, Johnson and Newport (1989) excluded a NS with accented speech, and Bongaerts et al. (1997) excluded speakers who made many errors and exhibited hesitations and pauses in their speech. Bongaerts and colleagues' decision may have been inspired by a previous experience in which the NS controls were found to perform relatively poorly (Bongaerts, Planken, & Schils, 1995), a finding that was attributed to the fact that the NSs of English spoke with a slight regional accent, whereas the Dutch learners were trained to use Received Pronunciation.

Of all studies reviewed and listed in Table 1, Coppieters (1987) is the only study that discussed the NSs at length, acknowledging that NSs may exhibit considerable variation in performance. Coppieters noted that "any study pretending to compare and contrast NS's [*sic*] and NNS's [*sic*] will have to face the problem of NS variability" (p. 548), and he acted accordingly by trying to select a sample of NSs of French that might be representative of the entire L1 community. He selected speakers from different regions of France and Belgium and tried to include people of lower and higher educational backgrounds. In a footnote to his replication of Coppieters's study, Birdsong (1992) concurred that "the putative 'ultimate attainment' of one group of native speakers may be different from that of another group of natives" (p. 707), but he also pointed out that in practice, most studies selected NS controls "that are similar along relevant social and educational dimensions" (p. 707). Coppieters's study also demonstrates the difficulties of constructing a good comparison group. Most important, it requires decisions about who can provide a fair NS norm. For Coppieters, this included Belgian speakers of French but not Canadian or Swiss speakers.

Table 1. List of CPH studies that compared NS and NNS performance

Study	L1-L2	L2 group	L1 norm group
Abrahamsson & Hyltenstam (2009)	Study 1 Spanish-Swedish	195 NNSs (AoO 1–7).	20 NSs: 10 from Stockholm and 10 with minor regional accents; no further information.
	Study 2 Spanish-Swedish	41 NNSs: 31 early (AoO ≤ 11) and 10 late (AoO > 12) learners.	15 NSs, matched with NNS sample on sex, educational level (at least senior high school), variety of Swedish, and age.
Asher & García (1969)	Spanish-English	71 NNSs: 56 early (AoO ≤ 11) and 15 late (AoO > 12) learners.	30 NSs, children; no further information. 19 NS judges, high school students.
Bialystok & Miller (1999)	Chinese-English and Spanish-English	Chinese: 33 NNSs: 15 early (AoO ≤ 15) and 18 late (AoO > 15) learners; Spanish: 28 NNSs: 15 early (AoO ≤ 15) and 13 late (AoO > 15).	38 NSs, no background information.
Birdsong (1992)	English-French	20 NNSs, all late learners.	20 NSs, all college-educated speakers of standard French. Note: It was explicitly acknowledged that NSs vary, but no measures were taken toward representativeness.
Birdsong (2007)	French-English	22 NNSs, all late learners.	17 NSs, all university educated. 3 NS judges, experienced teachers.
Birdsong & Molis (2001)	Spanish-English	71 NNSs: 29 early (AoO ≤ 16) and 32 late (AoO > 16) learners.	Provided by Johnson & Newport (1989): 23 NSs, no background information.
Bongaerts (1999)	Study 3 French-Dutch	9 NNSs, all late learners (AoO > 11).	9 NSs, selected for neutral accent; no further information. 10 NS judges.

Continued

Table 1. *Continued*

Study	L1-L2	L2 group	L1 norm group
Bongaerts, Mennen, & Van der Slik (2000)	Various-Dutch	30 NNSs, all late learners (AoO > 11).	10 NSs, in or graduated from tertiary education. 21 NS judges: 11 Dutch L2 teachers and 10 inexperienced judges. Note: 1 NS was excluded because of poor performance.
Bongaerts, Planken, & Schils (1995)	Dutch-English	10 NNSs, all late learners (AoO > 11).	5 NSs, no accents, university backgrounds. 4 NS judges, inexperienced, various occupations. Note: NSs were found to perform rather poorly, probably due to accent.
Bongaerts, Van Summeren, Planken, & Schils (1997)	Dutch-English	11 NNSs, all late learners (AoO > 11), and 20 randomly selected NNSs.	10 NSs, selected for speaking without an accent; no further information. 13 NS judges, selected for speaking without an accent. Note: NSs with accented speech or with many hesitations and errors in speech were excluded.
Colantoni & Steele (2006)	English-Spanish	10 NNSs, all late learners (AoO > 10).	10 NSs, no background information. Note: 1 NS was removed; unclear why.
Coppieters (1987)	Dutch-French	21 NNSs, all late learners (AoO > 17).	20 NSs: 10 were related to the NNSs, and the others were "representatives of various regions of France, as well as Belgium" (p. 551); varied educational backgrounds.

Flege et al. (2006)	Korean-English	36 NNSs, children (AoO ≤ 13). 36 NNSs, adults (AoO > 20).	18 NSs, children. 18 NSs, adults. Matched with NNS samples on age, English-speaking parents, no further information.
Flege & Liu (2001)	Chinese-English	60 NNSs, all late learners (AoO > 15).	5 NSs, no background information.
Flege, Munro, & MacKay (1995)	Italian-English	240 NNSs (AoO 2–23).	24 NSs, no background information.
Flege, Yeni-Komshian, & Liu (1999)	Korean-English	240 NNSs (AoO 1–23).	24 NSs, no background information.
Guion (2003)	Quechua-Spanish	20 NNSs: 15 early (AoO ≤ 14) and 5 late (AoO > 14) learners.	5 NSs, no background information. Note: Differences were based on acoustic analysis.
Guion, Harada, & Clark (2004)	Spanish-English	20 NNSs: 10 early (AoO < 7) and 10 late (AoO > 14) learners.	10 NSs, no background information.
Ioup, Boustagui, El Tigi, & Moselle (1994)	English-Arabic	1 NNS, late learner.	3–11 NSs, varied from task to task; all university graduates.
Johnson & Newport (1989)	Korean-English and Chinese-English	46 NNSs: 23 early (AoO ≤ 14) and 23 late (AoO > 17) learners.	23 NSs, no background information. Note: 2 NSs were excluded, one because the L1 was acquired outside the U.S. and the other because of a regional accent.
B. Lee, Guion, & Harada (2006)	Korean-English and Japanese-English	40 NNSs: 10 early Korean learners (AoO < 6) and 10 late Korean learners (AoO >15); 10 early Japanese learners (AoO > 15) and 10 late Japanese learners (AoO > 15).	10 adult NSs, monolingual, normal hearing; no further information.

Continued

Table 1. *Continued*

Study	L1-L2	L2 group	L1 norm group
D. Lee & Schachter (1997)	Korean-English	76 NNSs: approx. 48 early (AoO < 12) and 28 late (AoO ≥ 12) learners.	12 adult NSs, no background information.
McDonald (2000)	Spanish-English and Vietnamese-English	Spanish: 28 NNSs: 14 early (AoO < 6) and 14 late (AoO > 14) acquirers; Vietnamese: 28 NNSs: 14 early (AoO < 6) and 14 later (AoO < 10) acquirers.	14 NSs, no background information.
McDonald (2006)	Various-English	50 NNSs, all late (AoO > 11) learners.	50 NSs, university graduates.
Montrul & Slabakova (2003)	Spanish-English	64 NNS, all late learners.	20 NSs: from Spain (2), Argentina (12), Colombia (3), Costa Rica (2), and Mexico (1); 9 were tested in Argentina; 11 were tested in the U.S. No further information.
Moyer (1999)	English-German	24 late learners.	4 NSs, no background information.
Munro & Mann (2005)	(Mandarin) Chinese-English	32 NNSs (AoO 3–16).	4 NSs, university students, no regional or dialectal accent; no further information. 14 NS judges, university students.
Neufeld (1988)	French-English	22 advanced learners, 16 beginners.	12 NSs, limited exposure to English, enrolled at Faculty of Arts, 18–23 years old, born in the area, no knowledge of other languages.

Neufeld (2001)	English-French	18 late learners; 7 were identified as highly advanced.	3 NSs, university students; 2 of Canadian origin, 1 of French origin; limited knowledge of English. 68 NS judges, mostly students; limited knowledge of English; similar dialectal background, chosen from an area other than the area of study, because of the "broad phonological spectrum of French dialects" (p. 190) present in the area of study.
Oyama (1976)	Italian-English	60 NNSs (AoO 6–20); men only, all had college-level educational backgrounds.	No NS controls. 2 NS judges, students of linguistics.
Patkowski (1980)	Various-English	67 NNSs: 33 early (AoO ≤ 15) and 34 late (AoO > 15) learners.	15 NSs, highly educated. 2 NS judges, ESL teachers.
Tsukada, Birdsong, Bialystok, Mack, Sung, & Flege (2005)	Korean-English	24 NNSs: 12 early (AoO *M* = 10.5) and 12 late (AoO *M* = 15.6) learners.	No NSs involved in testing the CPH.
Van Boxtel, Bongaerts, & Coppen (2005)	Various-Dutch	43 late learners.	44 NSs, high level of education (undergraduate). Note: 2 NSs were excluded because of deviant scores, and 1 because he/she lived outside the Netherlands at a young age.
White & Genesee (1996)	Various-English	89 NNSs: 40 early (AoO > 12) and 49 late (AoO ≥ 12) learners.	19 NSs, no background information.

Note. AoO = age of onset.

The practice of NS comparison group sampling as it has been employed up until now is valid only if NSs exhibit very little variation. Until now, NS variation has not been a popular topic of investigation (Dąbrowska, 1997). The reason for this may be rooted in Chomsky's (1965) proposal that the competence to be studied in linguistics is that of the ideal speaker-listener "who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance" (p. 3). Although Chomsky acknowledges that sources of individual variation may be rooted in knowledge and the ability to apply that knowledge, many linguists have taken this to mean that NSs have uniform mental grammars and that individual differences have nonlinguistic—and therefore, theoretically uninteresting—causes only (Dąbrowska & Street, 2006). As stated by Dąbrowska (2012), the assumption of uniform success in L1 acquisition actually forms the basis for claims about a critical period and the fundamental difference between L1 and L2 acquisition, as it has been contrasted with the general failure of L2 learners to become nativelike.

However, insofar as evidence has been accumulated, it has been shown that NSs also vary considerably, even in the domain of grammar. Dąbrowska (2012) recently reviewed a number of studies on NS variation, some of which explicitly challenged the notion of a uniform NS competence, and the results supported the notion of experience-based individual differences (Dąbrowska, 1997; Dąbrowska, 2008; Dąbrowska & Street, 2006; Ferreira, 2003; Street & Dąbrowska, 2010). The common denominator in these studies was the testing of heterogeneous samples of adult NSs (in terms of educational level) on their knowledge of fairly common, but grammatically complex, structures such as passives (Dąbrowska & Street 2006) or the Polish genitive inflection (Dąbrowska 2008). The studies all attest to considerable differences between NSs in knowledge of the constructions under investigation. Dąbrowska (2012) suggests that the results are probably caused by differences in quantitative or qualitative experiences with processing linguistic texts, due to the participants' different educational backgrounds. In other words, NSs exhibit variation due to differences in quality and quantity of the input—that is, for the same reasons that NNSs do.

In light of such variation, the standard practice of NS norm group sampling in CPH research may be a cause for concern. In this study, several tasks with high- and low-frequency items were administered to 113 fairly advanced L2 learners as well as a heterogeneous sample of 120 NSs to compare the results of the NNSs against two different samples of NSs that differed depending on educational background. The data that are presented in this study were not gathered with the intention of falsifying the CPH and are not fit to do so for some of the reasons

discussed in this introduction: The L2 learners were probably not all highly advanced L2 users, and the tasks were not designed to test the CPH. The goal of the present study was to illustrate that the use of nonrepresentative NS norm groups affects the observed incidence of NNSs falling within the NS range. Additionally, I investigated how the use of basic versus more peripheral task items (in terms of frequency) affects these incidence rates. The expectation was that incidence rates drop as the educational level of the norm group increases and that incidence rates drop when low-frequency test items are used to provide the NS range.

METHOD

Participants

Native and nonnative speakers of Dutch were compared in this study; they were recruited in and around the city of Amsterdam through advertisements posted in several educational institutes, supermarkets, and community centers and through networks of relatives and friends. Selection was based on level of education (LoE; vocational or higher) and age (approximately between 20 and 35). The NS sample consisted of 124 participants (85 females, 39 males; 63 of low LoE, 61 of high LoE) ranging from 19 to 40 years old (M = 25, SD = 5.2). The low LoE participants were actively enrolled in vocational education or had already completed their studies, and some of the occupations present in the sample were house painter, gardener, mechanic, secretary, and train driver. The high LoE participants were actively enrolled in or had completed education at the bachelor's level (27 participants) or master's level (34 participants); in two cases these were language related. The high LoE participants held jobs such as managers, financial analysts, pedagogy counselors, and journalists. Although the NS group was rather heterogeneous in terms of occupations and educational backgrounds, the sample cannot be considered an accurate reflection of Dutch society. The lowest types of education were still underrepresented, and the sample was not randomly drawn.

The NNS sample consisted of 118 participants (80 females, 38 males; 52 of low LoE, 66 of high LoE) ranging from 19 to 40 years old (M = 29, SD = 5.3). Twelve participants were studying (or had studied) in a language program. As described before, the participants were not selected to be nativelike in Dutch; they varied in Dutch proficiency from B1 on the Common European Framework of Reference (Council of Europe, 2001) scales to C2, which means that they ranged from intermediate levels to nativelike levels of proficiency. Participants were asked to judge their own proficiency in reading, writing, listening, speaking,

grammar, and vocabulary on a 5-point Likert scale, yielding a summed maximum score of 30. The mean score was 21.5 (*SD* = 4.7). Seven participants gave themselves 30 out of 30 points and rated themselves as nativelike. The mean age at arrival in the Netherlands ranged from 0 to 36 years old (*M* = 21; *SD* = 8.1), and length of stay in the Netherlands varied from 8 months to 27 years (*M* = 8 years; *SD* = 6.1 years). Thirty-five different L1s were present in the NNS sample, which included German (9 speakers), Russian (9 speakers), Bahasa Indonesian (9 speakers), and Spanish (8 speakers).

All participants were financially compensated. Participants with hearing problems and people using medication that might impair their ability to perform reaction-time tasks were excluded from the study. All participants signed a consent form.

Tasks and Materials

Participants performed a range of tasks in a fixed order of administration; these included tasks not relevant for the present study, such as measures of working memory and intelligence. The tasks were mostly administered in two sessions of 2 hr on separate days with short breaks between the tasks, although some participants did all tasks in 1 day, with a long break between the two sessions. All tasks were newly developed, and their quality was assessed on the basis of pilots with 27–51 NSs and NNSs (this varied per task). Another goal was to make sure that the tasks were suitable for both groups. None of the tasks were designed with falsification of the CPH in mind, although vocabulary and grammaticality judgment tasks have regularly been used in CPH research. All tasks are also described in Andringa, Olsthoorn, Van Beuningen, Schoonen, and Hulstijn (2012) and Olsthoorn, Andringa, and Hulstijn (2012).

Segmentation (Word Recognition). The segmentation task was designed to test to what extent participants were able to recognize common words in normal speech. The test consisted of two parts. In the first part, participants heard short fragments of speech, and they were instructed to indicate as quickly and accurately as possible how many words they had heard by pressing the corresponding number on the keyboard. The test consisted of five practice trials and 45 experimental trials presented in random order and with 2 s intervals. The speech fragments consisted of two, three, and four words, but participants could press the numbers 1 to 5. In the second part of the test (not reported), participants listened to the same trials again and had to identify which words they heard by typing them out. Twenty-two fragments were fully articulated, and the articulation of the other 23 was reduced according to the principles of

vowel reduction or consonant reduction (or both) as described in Ernestus (2000), Kloots, De Schutter, Gillis, and Swerts (2003), and Coussé, Gillis, and Kloots (2007). These were typical of normal colloquial Dutch (compare *He has not* and *He hasn't* in English). Fragments always consisted of highly common words only: If the task was difficult, it was due to the reductions. All fragments were spoken by a female NS of Dutch with a neutral accent who was recorded in a soundproof studio; the recording was digitized at 44 kHz (16-bit quantization). The task was set up and run with the E-Prime software package (Schneider, Eschman, & Zuccolotto, 2002). Both speed (response time from fragment offset) and accuracy (number of trials correctly answered) were logged.

Grammatical Processing. The grammatical processing task assessed participants' knowledge of word-order properties of the Dutch language. Participants heard speech fragments of three to four words long and had to indicate whether the string was permissible at the beginning of a sentence by pressing *yes* or *no* as fast as possible. The task consisted of four practice trials and 34 experimental trials that were presented in random order and with 2 s intervals. Half of the stimuli constituted grammatically correct sentence beginnings, and the other half of the stimuli were ungrammatical and thus cannot occur in sentence-initial position. For instance, stimuli such as *Ze hebben overal* "They have everywhere" and *Die stad lijkt heel* "That city seems very" required a *yes* response, whereas *Precies ik weet* "Exactly I know" and *Er altijd moet iemand* "There always must somebody" required a *no* response. The stimuli were spoken at a normal pace by a female NS of Dutch with a neutral accent who was recorded in a soundproof studio; the recording was digitized at 44 kHz (16-bit quantization). The task was set up and run with the E-Prime software package (Schneider et al., 2002). Both response accuracy and response speed (delay between stimulus onset and response) were logged.

Semantic Processing. The semantic processing task was aimed at assessing the ability to comprehend the semantic and pragmatic meaning of full sentences. The task contained 60 trials: four practice trials and 56 test trials. Each trial consisted of a spoken proposition and two short response alternatives shown on the computer screen; one was a semantically and pragmatically meaningful response to the spoken proposition, and one was an inappropriate response. The participants' task was to indicate as quickly as possible which was the appropriate response by pressing the associated key. The response alternatives were visible on screen for 3 s before onset of the spoken proposition and remained visible until a decision was made for either. The interval between a response and the presentation of the next stimulus was 2 s. The propositions were spoken at a normal pace by a female NS of Dutch

with a neutral accent who was recorded in a soundproof studio; the recording was digitized at 44 kHz (16-bit quantization). The task was set up and run with the E-Prime software package (Schneider et al., 2002). Both response accuracy and speed were recorded. Response speed was computed as the interval between proposition onset and response-key press.

The spoken propositions varied along three parameters: length (between 7 and 12 words or between 13 and 23 words), syntactic complexity (with or without a subordinate clause), and content word frequency. In this study, only analyses pertaining to frequency will be presented: The high- and low-frequency stimuli were balanced in length and syntactic complexity. To determine frequency, the three least frequent content words of the sentence were identified in the CELEX corpus (Baayen, Piepenbrock, & Gulikers, 1995). The CELEX database is a corpus of written Dutch that consists of 124,000 lemmas of Dutch, with frequency information based on a corpus of written Dutch of 50 million word forms. A sentence was highly frequent if its three least frequent words occurred an average of more than 30 times per million words. For low- frequency sentences, the mean frequency of the content words had to be lower than 30 times per million. The response alternatives that participants chose from always consisted of highly common phrases, maximally four words long, as can be seen in the following examples:

(1) Proposition (short, simple, frequent):
Gaat de bus naar Amsterdam nog wel?
"Is the bus to Amsterdam still in service?"
Reactions:
Nee sorry / *Goed idee*
"No sorry" / "Good idea"

(2) Proposition (short, simple, infrequent):
Uit die reactie beluister ik onwil om mee te werken.
"In this response I detect reluctance to cooperate."
Reactions:
Nee sorry / *Ik ook*
"No sorry" / "I do too"

Word Monitoring. This task tested participants' ability to use their knowledge of the distributional and combinatorial properties of the Dutch language to predict upcoming information. In each trial, a target word to be monitored appeared on the screen for 1 s, and participants were instructed to remember that word. Subsequently, a carrier sentence was played, and participants were instructed to press the space bar as soon as they heard the target word. The monitor task consisted of 46 experimental trials and 4 practice trials, and they were presented in random order with 2 s intervals between each trial. The sentences

varied in length from 7 to 17 words; sometimes they consisted of main clauses only and other times of main clauses with coordinate, subordinate, or relative clauses. The position of the word to be monitored in the carrier sentence varied from 2nd to 14th position. The words to be monitored were nouns (n = 23), verbs (n = 8), adjectives (n = 6), prepositions (n = 5), and adverbs (n = 4). All carrier sentences consisted of simple, high-frequency words. The propositions were spoken at a normal pace by a female NS of Dutch with a neutral accent who was recorded in a soundproof studio; the recording was digitized at 44 kHz (16-bit quantization). The task was set up and run with the E-Prime software package (Schneider et al., 2002). Response time was logged from the onset of the word to be monitored. Participants could press before the word actually appeared, which means that negative values could and did occur.

Self-Paced Listening. The self-paced listening task was a measure of lower level sentence processing efficiency. The task was set up in accordance with the specifications provided by Ferreira, Henderson, Anes, Weeks, and McFarlane (1996). Participants listened to sentences one word at a time by pressing the space bar to hear the next word. If the space bar was pressed before the end of the word, it was truncated, and the next word was played. A third of the sentences were followed by a simple *yes-no* comprehension question to encourage participants to execute the task faithfully. Sentences were presented in random order. Participants were instructed to be as fast and accurate as possible. The task consisted of 56 trials, and participants were given four trials to practice. The sentences were spoken at a normal pace by a female NS of Dutch with a neutral accent who was recorded in a soundproof studio; the recording was digitized at 44 kHz (16-bit quantization). The task was set up and run with the E-Prime software package (Schneider et al., 2002). Response times were measured from the offset of each fragment within a sentence, and for each sentence, an average response time was calculated. Responses to sentence-initial and sentence-final fragments were removed because these tend to be highly variable.

The stimulus materials consisted of 56 grammatical sentences of Dutch. As in the semantic processing task, sentences varied in length, complexity, and frequency, and these parameters were operationalized according to the same criteria. The utterances were read word by word by a female NS of Dutch with a neutral accent who was recorded in a soundproof studio; the recording was digitized at 44 kHz (16-bit quantization). Although sentences were presented and recorded word by word, care was taken to retain the Dutch intonation contours as much as possible.

Receptive Vocabulary Knowledge. The vocabulary test was a computer-administrated, receptive multiple-choice test that measured passive knowledge of vocabulary. The test was based on a selection of items

from Hazenberg and Hulstijn (1996), a test originally developed for speakers of Dutch as a L2. Target vocabulary words were presented in a carrier sentence from which the meaning of the target word could not be deduced. Participants had five options to choose from, the last one always being "I really don't know." New low-frequency items were added to make the test suitable for NSs. They were constructed according to the same principles as those employed by Hazenberg and Hulstijn. First, target words were not jargon and did not represent specialized fields of knowledge. Also, items were created so that it should have been possible to capture the meaning of the target words in simple, high-frequency language and so that it was possible to present the target word in a simple carrier sentence. Care was taken not to introduce any systematicity in the length of the options and the way in which meanings were described.

The test consisted of 60 items. The target words were selected on the basis of frequency information from CELEX (Baayen et al., 1995), and they gradually decreased in frequency. For the purpose of this study, frequency was operationalized as a two-level factor by separating the 30 highest frequency items (once or more per million/42 occurrences or more in the total corpus of 50 millions words) from the 30 lowest frequency items (once or less per million/40 or fewer occurrences in the total corpus). Examples of high-frequency items are *glanzen* "to shine," *traject* "trajectory," and *toelage* "allowance"; examples of low-frequency items are *oploop* "gathering," "stir," *affreus* "hideous," and *hagiografie* "hagiography."

Statistical Analyses

Following a commonly used procedure in CPH research, I ascertained the incidence of NNSs falling within the NS range of scores. In previous studies, two procedures have been used to determine the NS range. One is simply the use of the observed statistical range, which means that the NNS score should be between the lowest and the highest NS score. Alternatively, the range has also been based on the mean and the standard deviation, whereby a NNS score should fall within a particular number of NS standard deviations from the NS mean to count as nativelike. Some have preferred the use of the statistical range (based on minimum and maximum scores), as it tends to be more inclusive (Abrahamsson & Hyltenstam, 2009), but this measure has two disadvantages. First, one person—the slowest or the lowest scoring NS participant—ultimately sets the standard. Good performance on a task is hardly ever the result of coincidence, but poor performance may well be due to factors that are irrelevant to the measurement, such as

attention lapses caused by any number of personal reasons or an off day, among others. Such factors should not determine the threshold. Additionally, the statistical range is sensitive to sample size: It tends to increase as the sample size grows. For these reasons, I adopted a range based on the mean score and the standard deviation. Also, I removed and replaced clearly outlying responses to individual trials. The threshold value beyond which I considered a NNS score within the NS range was the mean plus (for speed tasks) or minus (for accuracy tasks) 2 NS standard deviations. However, if this value exceeded the minimum or maximum observed value, it was replaced by that value, so that the threshold never fell outside the range of observed NS values. The obtained threshold was always rounded upward for accuracy tasks and downward for speed tasks: If the accuracy mean minus 2 standard deviations was 26.3, then scores of 27 and higher were considered within the NS range.

The first step in the analyses, then, was to inspect the scores for outliers. For the segmentation task, one NS was removed because he or she scored more than 40% incorrect on the fully articulated items. This was taken to mean that this person had not understood the task: In a second run of the task, participants were asked to type in the exact words heard, and no mistakes were made by this person on the fully articulated items. Nonnative speakers were also removed occasionally because of clearly outlying responses or because their data were not correctly logged (five NNSs were removed from the segmentation results, four from the grammatical processing results, two from the semantic processing results, and two from the word monitoring results). For the speed tasks (segmentation speed, grammatical processing speed, semantic processing speed, self-paced listening, and word monitoring), outliers on individual trials were identified. Responses were considered outliers if they fell outside the range of 2.5 standard deviations from the group mean for that item (this was done separately for NSs and NNSs), in which case they were coded as missing. This never constituted more than 2.4% of the data points in any task. Additionally, latencies of inaccurate responses were considered invalid and were also set to missing for segmentation speed, grammatical processing speed, and semantic processing speed (for the NSs: 12.8%, 3.6%, and 1.7%, respectively, and for the NNSs: 16.5%, 13.1%, and 6.7%). All missing data points were then newly estimated by imputation in SPSS by means of the full information maximum likelihood estimation procedure, and mean scores per participant were calculated.

To investigate how the NS sample affects the observed rate of NNSs falling within the NS range, representative and nonrepresentative NS norm groups were needed. For the nonrepresentative sample (NRS) norm, all master's-level graduates in the NS sample were selected, totaling 34 participants. For the representative sample (RS) norm, I attempted

to construct a sample that would be a fair reflection of the spread of educational levels in Dutch society. According to data gathered by the Centraal Bureau voor de Statistiek (CBS; 2011), 9% of the Dutch labor force possessed a master's-level degree in 2011; 18% possessed a bachelor's-level degree, and 73% possessed a vocational degree or no degree (the CBS distinguishes as many as seven subcategories within the vocational level, but these were not identifiable in the data). Thus, the educational backgrounds in the representative sample were equally proportioned and consisted of all lower level participants (58 participants), supplemented with 13 randomly drawn bachelor's-level participants (from a total of 27) and 6 randomly drawn master's-level participants (from a total of 34). Consequently, the representative sample norms were based on 77 participants. The data from master's-level participants were present in both the RS and NRS norms.

The two NS samples were compared for mean and standard deviation differences by means of one-way ANOVAs and Levene's test for equality of variances: Means and standard deviations determined the actual span of the RS and NRS ranges against which NNSs were tested. For the NNSs, I simply determined for each task whether their score fell inside or outside the NS RS and NRS ranges, which resulted in two nominal variables (inside or outside the RS and NRS) for each score. I ran exact McNemar tests and a chi-square test for related pairs (there were two decisions for each participant: one based on the RS norm and one based on the NRS norm) to assess whether the NRS norm excluded significantly more NNSs as nativelike than did the RS norm, which was the hypothesis. Figure 1 illustrates how the McNemar test works. It determines whether the probability of *b* is equal to the probability of *c*, in which the letters refer to observed cell frequencies. Thus, the test considers only those observations that fall inside the RS range and outside

		Nonrepresentative sample (NRS) norm		
		Score outside norm	Score inside norm	Total
Representative sample (RS) norm	Score outside norm	*a*	*b*	*a* + *b*
	Score inside norm	*c*	*d*	*c* + *d*
	Total	*a* + *c*	*b* + *d*	*n*

Figure 1. Example contingency table.

the NRS range and vice versa; it ignores those observations that are included or excluded by both norms.

RESULTS

Table 2 displays the descriptive statistics for the RS and the NRS NS norm groups on all measures. One-way ANOVAs were used to test for differences between the two groups. The table shows that significant differences in mean scores were observed for the four accuracy tasks. The NRS group was better on these tasks. For the speed tasks, no mean differences were observed, with the exception of segmentation speed, in which the NRS group was faster. Additionally, differences in standard deviations were observed for grammatical processing accuracy, segmentation speed, and word monitoring as indicated by violations of Levene's test for equality of variances. In all cases, the NRS group displayed less variation in performance. For semantic processing speed, grammatical processing speed, and self-paced listening, no differences were observed in either mean scores or variances.

To test whether the incidence rates were affected by the nature of the NS norm group, the range of performances of the NNS sample was compared with a RS and a NRS NS norm. For each task, I determined how many NNSs scores fell inside or outside the RS and NRS range (Table 3). For vocabulary, Table 3 shows that 63 L2 learners were excluded by both norms, and 29 learners were included by both. However, 22 learners (19%) fell outside the NRS norm but inside the RS norm. This suggests that it is easier to meet the norm set by the RS than the norm set by the NRS. Although the numbers differed from task to task, this pattern of results was obtained for all of the tasks. McNemar tests were performed to assess whether the observed distributions may be considered significantly different from one another. This was true for all measures, even for those measures for which I did not observe significant mean or variance differences.

The second hypothesis predicted that fewer L2 learners would fall within the NS norm if the norm were based on the low-frequency task items only. For these analyses, only the results for vocabulary, semantic processing accuracy and speed, and self-paced listening were included in the analyses, as they were the only tasks that were manipulated for frequency. Table 4 shows the descriptive statistics for NSs, split out by norm group and frequency.

Table 5 shows whether the L2 learners fell inside or outside the NS norm as set by low-frequency and high-frequency task items. For vocabulary, 25 learners fell outside and 35 fell inside the RS NS range, irrespective of whether the range was calculated on the basis of high-frequency items only or low-frequency items only. Two learners fell

Table 2. Descriptive statistics, threshold values, and mean and standard deviation comparisons for the representative sample (RS) and nonrepresentative sample (NRS) native speaker groups

Task	Group	*N*	*M* (*SD*)	Min–max	Threshold value (*M* +/– 2 *SD*s)	*M* diff. *p* value	*SD* diff. *p* value
Vocabulary (max score = 60)	RS	77	39.4 (5.49)	27–54	29	.00	.73
	NRS	34	44.7 (5.49)	35–60	35		
Semantic processing accuracy	RS	75	54.9 (1.13)	51–56	53	.01	.17
(max score = 56)	NRS	33	55.5 (0.87)	52–56	54		
Grammatical processing accuracy	RS	76	32.3 (1.53)	27–34	30	.04	.03
(max score = 34)	NRS	34	32.9 (1.04)	31–34	31		
Segmentation accuracy	RS	76	40.5 (2.00)	36–44	37	.00	.07
(max score = 44)	NRS	34	41.9 (1.49)	38–44	39		
Semantic processing speed	RS	75	3,540 (608)	2,197–4,963	4,757	.12	.76
	NRS	33	3,349 (536)	2,365–4,264	4,264		
Grammatical processing speed	RS	76	1,505 (256)	904–2,081	2,017	.42	.14
	NRS	34	1,464 (211)	1,096–1,966	1,887		
Segmentation speed	RS	76	1,389 (430)	593–2,621	2,250	.03	.05
	NRS	34	1,213 (308)	435–2,132	1,831		
Word monitoring	RS	75	276 (67)	107–436	410	.60	.03
	NRS	34	282 (46)	210–420	375		
Self-paced listening	RS	72	1,613 (92)	1,380–1,872	1,798	.94	.29
	NRS	34	1,612 (73)	1,466–1,758	1,758		

Note. Min = minimum; max = maximum; diff. = difference.

Table 3. Observed incidence of NNSs falling inside or outside the representative sample (RS) and nonrepresentative sample (NRS) ranges based on the mean +/– 2 SDs

Task	*N*	*n* outside both ranges	*n* inside both ranges	*n* inside RS–outside NRS range	*n* outside RS–inside NRS range	McNemar *p* value
Vocabulary	114	63	29	22	0	.00
Semantic processing accuracy	109	43	50	16	0	.00
Grammatical accuracy	109	50	49	10	0	.00
Segmentation accuracy	113	79	23	11	0	.00
Semantic processing speed	107	38	53	16	0	.00
Grammatical processing speed	103	45	49	9	0	.00
Segmentation speed	110	39	47	24	0	.00
Word monitoring	111	37	63	11	0	.00
Self-paced listening	109	29	74	6	0	.03

inside the high-frequency range but outside the low-frequency range, whereas 52 learners fell outside the high-frequency range but inside the low-frequency range. The McNemar test confirmed that L2 learners were significantly more likely to fall within the NS range on the low-frequency items. The same result was found for the semantic processing speed task when the NNSs were compared against the NRS norm. All other McNemar *p* values were not significant. Thus, the second hypothesis was not confirmed. If anything, the opposite was true.

DISCUSSION

This study was conducted to investigate how the selection of NSs in CPH research may affect decisions about the level of nativelikeness of NNSs. The literature review has shown that many CPH researchers do not provide much information about their NS sample or have not sampled NSs from a range of academic and vocational backgrounds. In CPH research, significant effort has been invested in describing the NNS sample, whereas the NS sample has received little attention in terms of text devoted to describing the sample. As a result, for many studies reviewed, it was impossible to determine who the NSs were exactly. In studies that provided more detail, the NSs often turned out to be highly educated. The data presented here showed that means and standard

Table 4. Descriptive statistics, threshold values, and mean and standard deviation comparisons for the representative sample (RS) and nonrepresentative sample (NRS) native speaker groups, split by frequency

Task	Group	Frequency	*N*	Max possible	*M* (*SD*)	Observed min–max	Threshold (*M* +/– 2 *SD*s)
Vocabulary	RS	High	77	30	25.4 (2.61)	18–30	21
	RS	Low	77	30	14.0 (3.67)	5–26	7
	NRS	High	34	30	27.4 (1.79)	23–30	24
	NRS	Low	34	30	17.4 (4.13)	10–30	10
Semantic processing accuracy	RS	High	75	27	26.5 (.74)	24–27	26
	RS	Low	75	29	28.4 (.75)	26–29	27
	NRS	High	33	27	26.9 (.42)	25–27	27
	NRS	Low	33	29	28.6 (.61)	27–29	28
Semantic processing speed	RS	High	75	n.a.	3,110 (547)	1,962–4,323	4,204
	RS	Low	75	n.a.	3,940 (676)	2,417–5,558	5,293
	NRS	High	33	n.a.	2,945 (466)	2,044–3,707	3,707
	NRS	Low	33	n.a.	3,725 (612)	2,533–4,782	4,782
Self-paced listening	RS	High	72	n.a.	1,615 (92)	1,417–1,881	1,799
	RS	Low	72	n.a.	1,611 (94)	1,343–1,862	1,799
	NRS	High	34	n.a.	1,616 (77)	1,478–1,776	1,770
	NRS	Low	34	n.a.	1,608 (72)	1,454–1,741	1,741

Note. Max = maximum; min = minimum; n.a. = not applicable.

Table 5. Observed incidence of NNSs falling inside or outside the high-frequency (HF) and low-frequency (LF) ranges, split by norm group

Task	Group	*N*	*n* outside both ranges	*n* inside both ranges	*n* inside HF–outside LF range	*n* outside HF–inside LF range	McNemar *p* value
Vocabulary	RS	114	25	35	2	52	.00
	NRS	114	53	17	3	41	.00
Semantic processing accuracy	RS	109	19	54	23	11	.06
	NRS	109	45	26	15	21	.41
Semantic processing speed	RS	107	33	68	1	5	.22
	NRS	107	50	48	0	9	.00
Self-paced listening	RS	109	28	79	1	1	1.00
	NRS	109	33	69	6	1	.13

deviations on rather unchallenging measures of language comprehension tended to differ significantly between the nonrepresentative, highly educated sample and a more representative sample in terms of educational level. Native speaker means and standard deviations determine the range against which NNSs are compared, and therefore they affect the observed incidence rate of nativelikeness. In this study, I found consistent confirmation of the hypothesis that the use of a representative NS sample lowers the threshold of nativelike performance. This was true even for measures for which I did not find significant mean or standard deviation differences between the RS and the NRS.

Whereas the first hypothesis was confirmed, the second was not. Incidence rates were not lower when the range was based on low-frequency test items. In fact, insofar as significant effects were observed, they pointed to the opposite. Scores based on high-frequency materials tended to exclude more NNSs from the NS range than scores based on materials of lower frequency. This was probably caused by the fact that there tended to be less variance among the NSs in the scores based on high-frequency item sets. Less variance meant smaller standard deviations, which in turn led to a smaller range and a higher threshold. In the low-frequency item sets, the NSs varied much more and had larger standard deviations. Therefore, the NS range was much wider and the threshold that NNSs had to meet was consequently lower. The effect was strong in the vocabulary task, on which 50 out of 114 NNS participants fell outside the NS RS range based on the high-frequency items but inside the range based on low-frequency items. For this task as well as for the semantic processing speed task, no ceilings were observed. The effect was not observed for semantic processing accuracy, which is probably due to

the fact that NS ceiling performance was observed not only on the high-frequency items but also on the low-frequency items. It was also not found for self-paced listening; for this task, there were simply no differences in means and standard deviations between the RS and the NRS or between high- and low-frequency scores.

As stated in the introduction, whether a NNS score falls within the NS range of scores is not just the result of which NSs are selected. It is also determined by which aspects of proficiency are tested as well as by the characteristics of the NNS sample. Given that the NNSs were not selected to be near-native, the observed incidence of nativelikeness was rather high in this study: On several tasks, more than half of the NNSs were able to score within the NS range. This was caused entirely by the comprehension tests that were used. They were developed to gauge the listening comprehension abilities of both NSs and NNSs (Andringa et al., 2012; Olsthoorn et al., 2012). As can be judged from the examples given in the Tasks and Materials section, the lexical content of the lower frequency items was probably unproblematic for many relatively advanced L2 learners, with the exception of the vocabulary test, for which no ceilings were observed, not even for the NSs.

A few questions remain to be addressed. For example, would these results have been obtained if tests had been used that are more appropriate to test the claims of the CPH or if truly near-native learners had been selected? These questions are difficult to answer, but the effects of LoE are not unique to this study; they were also observed in many of the studies on grammatical competence reviewed by Dąbrowska (2012). Thus, it seems likely that they also appear when other language measures are used. One could argue that the LoE effect was due not to differences in language ability but to differences in the ability to deal with the demands of the task. However, given the simplicity of the tasks used in this study, claiming that these findings could be the result of task effects only is probably a bit far-fetched. Similarly, it seems reasonable to assume that NSs will perform more homogeneously on simple, high-frequency language materials, irrespective of what task is used or what domain of language is tested. Whether that will always result in lower incidence of nativelikeness rates for high-frequency materials is an intriguing question that needs to be tested empirically. It also remains to be seen whether this works with highly advanced learners.

Hyltenstam and Abrahamsson (2003) have made several recommendations concerning the design of CPH studies. Most important, they recommended that highly advanced learners be subjected to linguistic scrutiny that involves "large sets of elicitation techniques for varied aspects of proficiency" (p. 577). This approach safeguards against claims of nativelikeness on the basis of excellent performance in just one of the many domains of language proficiency. However, they added the recommendation that ceiling effects should be avoided for NSs

and NNSs. The results of this study suggest that ceiling effects in NS performance may actually be helpful in identifying nonnative performance. In fact, one might argue that ceiling performance is what constitutes nativelikeness. In CPH research, the objective has been to test whether L2 learners know and can do as much as NSs. However, by selecting highly educated NSs, we may often have been testing at levels of proficiency that not all NSs reach. This approach is illustrated in Abrahamsson and Hyltenstam (2009). They developed highly challenging tests of phonological, grammatical, and lexical ability, tests on which their sample of 15 NSs were not able to perform at maximum levels. The 50-item test assessing knowledge of proverbs—on which the NSs scored a mean of 39 correct answers, ranging from 33 to 46—turned out to be the most difficult for NNSs: No more than 15% (6 speakers) scored within the NS range. Lost in this comparison of total scores is information about who knew what. On average, 11 proverbs were not known by the NSs. Were these always the same proverbs? More important, were there proverbs that were consistently known by all NSs? A comparison of that set of proverbs might constitute a better test of nativelikeness; the results of the present study suggest that even fewer NNSs would fall within the NS range.

Perhaps Hulstijn's (2011) notions of basic and higher language cognition are helpful tools in solving the paradox outlined in the introduction of how many tests suffice to conclude that someone is or is not nativelike. A tentative answer to Birdsong's (2005) question of which areas of language to consider for falsification of the CPH might be those areas in which NSs do not vary: basic language cognition, which is a finite domain, according to Hulstijn. The results of the present study may be construed as support for Hulstijn's theory in that they show that there are domains of language cognition in which a fairly heterogeneous group of NSs perform fairly homogenously as well as domains that are characterized by much more variation. As already noted, Hulstijn's theory has not been empirically tested, but perhaps a practical approach can be adopted in the effort to falsify the CPH. This approach would copy most of the recommendations made by Hyltenstam and Abrahamsson (2003), including the recommendation of investigating nativelikeness on many different aspects of L2 proficiency and in domains that are expected to be problematic for the L2 sample (e.g., because of the differences between the L1 and the L2). Crucially, however, the tasks should be demanding for the NNSs only, not for the NSs. As suggested earlier, this can be achieved by eliminating those items on which NSs exhibit variable performance.

The results of the present study have shown that the incidence of NNSs falling within the NS range is affected by the selection of the NSs. Who, then, should be selected? Ultimately, this is a theoretical question, too, not an empirical one. As such, the results of this study do not provide criteria for the selection of NSs. The common practice seems to have

been to compare NNSs with NSs who are generally highly educated. This may be fair, but probably only if the NNSs are comparable to the NSs in terms of the linguistic contexts in which they function. This would suggest that NSs and the NNSs should be matched on relevant background characteristics, as was done in Abrahamsson and Hyltenstam (2009). However, this leaves room for the possibility that none of the L2 learners will be found to be fully nativelike, although many of them may actually be more proficient language users than many NSs are. Therefore, there is something to be said for Coppieters's (1987) approach: Select a NS sample that is truly representative of the L1 speech community (although this triggers the sociolinguistic problem of defining a speech community). The NS sample in the present study consisted of speakers from the Amsterdam area, including one or two second-generation immigrants. Should I have included Dutch-Frisian bilinguals and Flemish speakers of Dutch? Such questions are not easy to answer.

In a recent methodological treatment, DeKeyser (2013) acknowledged that NS variation has not been considered much in CPH research. He recommended the use of a homogeneous NS sample: "When there is substantial variation among native speakers, then the native and non-native ranges of variation are almost bound to overlap substantially, whether for the same reasons or not, yielding trivially predictable results" (p. 58). He also recommended that the L1 and the L2 be distant languages so that one could compare them on more central (rather than peripheral) structural features of the target language. DeKeyser's recommendations are well worth noting but also raise questions. The NS sample should be homogeneous, but which NSs should be selected? Should they be highly educated and highly experienced language users? When is overlap between NS and NNS performance trivial, and when is it meaningful? Which languages are sufficiently different, and which structures are sufficiently central? Ultimately, these questions cannot be answered on strictly empirical grounds. We still lack a sufficiently comprehensive theory of what it means to master a language as a NS. Without that, we cannot decide on what aspects of the language NNSs should be tested, and we cannot know which NSs should be selected as the norm group. When it comes to selecting the NS norm group, we have seen that the NS samples used in CPH research have been quite small, whereas the NNS samples are often quite large. It should be the other way around. Long (1990) has claimed that it may be enough to identify just one NNS who is nativelike in every respect to falsify the CPH. If that is true, then we do not need large NNS samples. What we really need is a sizeable and well-described NS norm group.

Received 21 December 2012
Accepted 4 April 2013
Final Version Received 26 July 2013

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59*, 249–306.

Andringa, S. J., Olsthoorn, N. M., Van Beuningen, C. G., Schoonen, R., & Hulstijn, J. H. (2012). Determinants of success in native and nonnative listening comprehension: An individual differences approach. *Language Learning*, *62*(s2), 28–48.

Asher, J., & García, G. (1969). The optimal age to learn a foreign language. *Modern Language Journal*, *38*, 334–341.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Philadelphia.

Bialystok, E. (2005). Consequences of bilingualism for cognitive development. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 417–432). Oxford: Oxford University Press.

Bialystok, E., & Miller, B. (1999). The problem of age in second language acquisition: Influences from language, structure, and task. *Bilingualism: Language and Cognition*, *2*, 127–145.

Birdsong, D. (1992). Ultimate attainment in second language acquisition. *Language*, *68*, 706–756.

Birdsong, D. (2005). Nativelikeness and nonnativelikeness in L2A research. *International Review of Applied Linguistics*, *43*, 319–328.

Birdsong, D. (2007). Nativelike pronunciation among late learners of French as a second language. In O. Bohn & M. J. Munro (Eds.), *Second language speech learning: The role of language experience in speech perception and production* (pp. 99–116). Amsterdam: Benjamins.

Birdsong, D., & Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of Memory and Language*, *44*, 235–249.

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning? In S. M. Gass & J. Schachter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 41–68). Cambridge: Cambridge University Press.

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 133–159). Mahwah, NJ: Lawrence Erlbaum.

Bongaerts, T., Mennen, S., & Van der Slik, F. (2000). Authenticity of pronunciation in naturalistic second language acquisition: The case of very advanced late learners of Dutch as a second language. *Studia Linguistica*, *54*, 298–308.

Bongaerts, T., Planken, B., & Schils, E. (1995). Can late learners attain a native accent in a foreign language? A test of the critical period hypothesis. In D. Singleton & Z. Lengyel (Eds.), *The age factor in second language acquisition* (pp. 30–50). Clevedon, UK: Multilingual Matters.

Bongaerts, T., Van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, *19*, 447–465.

Centraal Bureau voor de Statistiek. (2011). *Statline databank* [Online database]. Retrieved from http://www.cbs.nl

Chomsky, N. (1965). *Aspects of theory and syntax*. Cambridge, MA: MIT Press.

Colantoni, L., & Steele, J. (2006). Native-like attainment in the L2 acquisition of Spanish stop-liquid clusters. In C. A. Klee & T. L. Face (Eds.), *Selected proceedings of the 7th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages* (pp. 59–73). Somerville, MA: Cascadilla Proceedings Project.

Cook, V. J. (2002). Background to the L2 user. In V. J. Cook (Ed.), *Portraits of the L2 user* (pp. 1–28). Clevedon: Multilingual Matters.

Coppieters, R. (1987). Competence differences between native and near-native speakers. *Language*, *63*, 544–573.

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

Coussé, E., Gillis, S., & Kloots, H. (2007). Verkort, verdoft, verdwenen: Vocaalreductie in het Corpus Gesproken Nederlands [Shortened, reduced, disappeared: Vocal reduction in the Corpus of Spoken Dutch]. *Nederlandse Taalkunde*, *12*, 109–138.

Dąbrowska, E. (1997). The LAD goes to school: A cautionary tale for nativists. *Linguistics*, *35*, 735–766.
Dąbrowska, E. (2008). The later development of an early-emerging system: The curious case of the Polish genitive. *Linguistics*, *46*, 629–650.
Dąbrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, *2*, 219–253.
Dąbrowska, E., & Street, J. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and non-native English speakers. *Language Sciences*, *28*, 604–615.
DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.
DeKeyser, R. (2012). Age effects in second language learning. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 442–460). New York: Routledge.
DeKeyser, R. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning*, *63*(s1), 52–67.
Ernestus, M. T. C. (2000). *Voice assimilation and segment reduction in casual Dutch: A corpus-based study of the phonology-phonetics interface* (Doctoral dissertation). Utrecht, the Netherlands: LOT.
Eubank, L., & Gregg, K. R. (1999). Critical periods and (second) language acquisition: Divide et impera. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 65–99). Mahwah, NJ: Erlbaum.
Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164–203.
Ferreira, F., Henderson, J. M., Anes, M. D., Weeks, P. A., & McFarlane, D. K. (1996). Effects of lexical frequency and syntactic complexity in spoken-language processing: Evidence from the auditory moving-window technique. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 324–335.
Flege, J. E., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics*, *34*, 153–175.
Flege, J. E., & Liu, S. (2001). The effect of experience on adults' acquisition of a second language. *Studies in Second Language Acquisition*, *23*, 527–552.
Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, *97*, 3125–3134.
Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, *41*, 78–104.
Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *23*, 311–343.
Gregg, K. R. (1996). The logical and developmental problems of second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 49–81). San Diego: Academic Press.
Grosjean, F. (1998). Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition*, *1*, 131–149.
Guion, S. (2003). The vowel systems of Quichua-Spanish bilinguals: Age of acquisition effects on the mutual influence of the first and second languages. *Phonetica*, *60*, 98–128.
Guion, S., Harada, T., & Clark, J. J. (2004). Early and late Spanish-English bilinguals' acquisition of English word stress patterns. *Bilingualism: Language and Cognition*, *7*, 207–226.
Hazenberg, S., & Hulstijn, J. H. (1996). Defining a minimal receptive second-language vocabulary for nonnative university students: An empirical investigation. *Applied Linguistics*, *17*, 145–163.
Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*, 229–249.
Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, *15*, 422–433.

Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in SLA. In C. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 539–588). Oxford: Blackwell.

Ioup, G., Boustagui, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study in a naturalistic environment. *Studies in Second Language Acquisition*, *16*, 73–98.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.

Kloots, H., De Schutter, G., Gillis, S., & Swerts, M. (2003). Vocaalreductie in spontaan gesproken Standaardnederlands: Een verkennende studie [Vocal reduction in spontaneously spoken standard Dutch: An exploratory study]. In T. Koole, J. Nortier, & B. Tahitu (Eds.), *Artikelen van de vierde sociolinguistische conferentie* (pp. 224–233). Delft: Eburon.

Lee, B., Guion, S., & Harada, T. (2006). Acoustic analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals. *Studies in Second Language Acquisition*, *28*, 487–513.

Lee, D., & Schachter, J. (1997). Sensitive period effects in binding theory. *Language Acquisition*, *6*, 333–362.

Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.

Long, M. H. (1990). Maturational constraints on language development. *Studies in Second Language Acquisition*, *12*, 251–285.

Long, M. H. (2005). Problems with supposed counter-evidence to the critical period hypothesis. *International Review of Applied Linguistics in Language Teaching*, *43*, 287–317.

Long, M. H., & Robinson, P. (1998). Focus on form: Theory, research, and practice. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 16–41). Cambridge: Cambridge University Press.

McDonald, J. L. (2000). Grammaticality judgments in a second language: Influences of age of acquisition and native language. *Applied Psycholinguistics*, *21*, 395–423.

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanation for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, *55*, 381–401.

Montrul, S. A. (2008). *Incomplete acquisition in bilingualism*. Amsterdam: Benjamins.

Montrul, S. A., & Slabakova, R. (2003). Competence similarities between native and near-native speakers: An investigation of the preterite-imperfect contrast in Spanish. *Studies in Second Language Acquisition*, *25*, 351–398.

Moyer, A. (1999). Ultimate attainment in L2 phonology: The critical factors of age, motivation and instruction. *Studies in Second Language Acquisition*, *21*, 81–108.

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics*, *29*, 578–596.

Munro, M. J., & Mann, V. (2005). Age of immersion as a predictor of foreign accent. *Applied Psycholinguistics*, *26*, 311–341.

Neufeld, G. (1988). Phonological asymmetry in second-language learning and performance. *Language Learning*, *4*, 531–559.

Neufeld, G. (2001). Non-foreign-accented speech in adult second language learners: Does it exist and what does it signify? *ITL International Journal of Applied Linguistics*, *133–134*, 185–206.

Olsthoorn, N. M., Andringa, S. J., & Hulstijn, J. H. (2012). Visual and auditory digit-span performance in native and nonnative speakers. *International Journal of Bilingualism*. Advance online publication. doi: 10.1177/1367006912466314

Oyama, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, *5*, 261–285.

Patkowski, M. S. (1980). The sensitive period for the acquisition of syntax in a second language. *Language Learning*, *30*, 449–472.

Penfield, W., & Roberts, L. (1959). *Speech and brain mechanisms*. Princeton, NJ: Princeton University Press.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh, PA: Psychology Software Tools Inc.

Street, J., & Dąbrowska, E. (2010). More individual differences in language attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua*, *120*, 2080–2094.

Tsukada, K., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Flege, J. E. (2005). A developmental study of English vowel production and perception by native Korean adults and children. *Journal of Phonetics*, *33*, 263–290.

Van Boxtel, S., Bongaerts, T., & Coppen, P. (2005). Native-like attainment of dummy subjects in Dutch and the role of the L1. *International Review of Applied Linguistics in Language Teaching*, *43*, 355–380.

White, L., & Genesee, F. (1996). How native is near-native? The issue of ultimate attainment in adult second language acquisition. *Second Language Research*, *12*, 238–265.