Used to be a dime, now it's a dollar: R-SPiN keyword predictability revisited 40 years on

Alexina Whitley[1], Graham Naylor[1], Lauren V Hadley[1]

Corresponding author:

Alexina Whitley

e-mail: Alexina.Whitley@nottingham.ac.uk

School of Medicine,

University of Nottingham,

Hearing Sciences – Scottish Section,

New Lister Building Level 3, Glasgow Royal Infirmary,

10-16 Alexandra Parade, Glasgow G31 2ER

[1]Hearing Sciences - Scottish Section, University of Nottingham

Funding Sources

This work was supported by a UKRI Future Leaders Fellowship [grant number MR/T041471/1]; and the Medical Research Council [grant number MR/X003620/1].

Declarations of Interest

None.

Abstract

Purpose: Almost 40 years after its development, in this paper we re-examine the relevance and validity of the ubiquitously used Revised-Speech in Noise sentence corpus (R-SPiN). The R-SPiN corpus includes 'high context and 'low context' sentences, and has been widely used in the field of hearing research to examine the benefit derived from semantic context across English-speaking listeners, but research investigating age differences has yielded somewhat inconsistent findings. We assess the appropriateness of the corpus for use today in different English-language cultures (i.e., British, and American), as well as for younger and older adults.

Method: 120 participants, including older (60-80 years) and younger (19-31 years) adult groups in the US and UK completed a cloze task consisting of R-SPiN sentences with final word removed. Cloze, as a measure of predictability, and entropy, as a measure of response uncertainty, were compared between culture and age groups.

Results: Most critically, of the 200 'high context' stimuli, only around half were assessed as highly predictable for older adults (UK: 109; US: 107), and fewer still for younger adults (UK: 75; US: 81). We also found dominant responses to these 'high context' stimuli varied between cultures, with US responses being more likely to match the original R-SPiN target.

Conclusions: Our findings highlight the issue of incomplete transferability of corpus items across English-language cultures, as well as diminished equivalency for younger and older adults. By identifying relevant items for each population, this work could facilitate the interpretation of inconsistent findings in the literature, particularly relating to age effects.

Introduction

Speech Listening and Context

Listening to speech is an incredibly complex task that involves the smooth and rapid coordination of several auditory and cognitive processes. In everyday conversation, the listener is challenged to follow the talker's speech, extract the meaning of each utterance, store it in memory for later recall, and integrate the incoming information with their own world knowledge, in order to generate an appropriate response (Schneider, 2011;Schneider, Bruce, & Pichora-Fuller, 2010). The overall demands placed on the listener increase when background noise is present, due to the acoustic degradation of the target speech and the need to inhibit other auditory streams (Mattys et al., 2012; Johnsrude & Rodd, 2016). But while such challenging conditions have detrimental effects on comprehension, listeners have been found to readily adapt to degraded speech by using top-down predictive processes guided by contextual information (Obleser, 2014).

Indeed, the context in which we hear words being spoken substantially impacts how we process that speech. A variety of different forms of context are present in everyday conversation, from prior experience of a talker's voice (Johnsrude et al., 2013; Brungart et al., 2001), to semantic and syntactic information (Obleser et al., 2007; Obleser & Kotz, 2010, 2011), or concurrent visual cues (e.g., lip movements; Keil et al., 2012 or gestures; Drijvers & Özyürek, 2017; Holler & Levinson, 2019). Listeners can use these contextual cues to better identify degraded speech, as they provide additional constraints which support processing. For instance, listeners are better able to comprehend speech in noise when the sentence provides a rich semantic context (Van Engen et al., 2014;Golestani et al., 2009), or when it is accompanied by relevant facial expressions (McGettigan et al., 2012), or gestures (Obermeier et al., 2012). Even in the absence of speech degradations, such contextual constraints can speed

understanding and reduce listening effort (Altmann & Kamide, 1999; van der Feest et al., 2019; Lash et al., 2013; Winn, 2016).

Here, we focus on linguistic context (the frame built by the meaning of words and structure of sentences), and more specifically semantic context. Listeners readily use such context to make predictions about upcoming words (Federmeier, 2007; Kuperberg & Jaeger, 2016; Pickering & Gambi, 2018). To test whether semantic context is used predictively (as opposed to facilitating retrospective postdiction of missed words), studies in this genre typically use a sentence-final keyword paradigm. For example, in the sentence beginning with *the boy eats the*, the range of possible upcoming words is constrained by the verb *eats*. Indeed, in studies using the visual world eye-tracking paradigm, when participants hear the verb *eats*, they show anticipatory eye movements towards edible objects (over inedible objects) before the onset of the noun that identifies the object (Altmann & Kamide, 1999).Use of semantic prediction is further evidenced by findings from EEG studies. Final word predictability has been shown to modulate the N400, a negative deflection associated with meaning-based processing that peaks 400ms after stimulus onset, whereby unexpected words elicit greater N400 amplitudes than expected words (DeLong et al., 2005; Kutas & Federmeier, 2010).

A number of corpora have been created to assess the benefit of linguistic context (Block & Baldwin, 2010; Bloom & Fischler, 1980), and these typically include matched sentences that differ in contextual constraint while being similar in other respects (e.g., word frequencies, phonetic balance). One key corpus is the Revised Speech Perception in Noise sentence corpus (R-SPiN; Bilger et al., 1984), whose influence is reflected in over 700 citations of the source publication. In this paper, we investigate how well the linguistic context manipulation in the R-SPiN corpus works today, almost 40 years after it was developed, across participant age groups and across two English-language cultures.

Importantly, we note that the term ‘context’ can encompass use of either prior or subsequent information. While the former would involve using information that came before a target word to aid its

processing, the latter would involve using information that came afterwards. Given the structure of the R-SPiN, in the present study we are concerned only with use of prior context, and hence we discuss stimuli in terms of 'predictability' and its etymological derivatives. However, in order to distinguish between the class originally assigned to a sentence by the R-SPiN developers, and the class our data suggest for that sentence, we label the former as 'High Context' ('HC') or 'Low Context' ('LC'), in inverted commas, and the latter as High Predictability (HP) or Low Predictability (LP), without inverted commas. References to 'high/low cloze words' are understood to mean sentence-final words whose predictability based on their preceding sentence frame is high/low, as derived from sentence-completion studies.

Brief overview of R-SPiN

The R-SPiN corpus consists of 400 phonetically balanced sentences with sentence-final monosyllabic target words, comprising 200 matched pairs of 'high context' ('HC') and 'low context' ('LC') sentences. These sentences have identical target words, but an 'HC' sentence provides extensive semantic cues to the target word's identity whereas an 'LC' sentence provides little or none. For example, the target *SPOON* occurs in the 'HC' sentence *He stirs the coffee with a SPOON,* and in the 'LC' sentence *Bob could have known about the SPOON*. Sentences are divided into eight lists of 50 (25 'HC' and 25 'LC'), with paired 'HC' and 'LC' sentences never occurring in the same list. Each odd numbered list has a matching even numbered list in which target words are presented in the opposite context (i.e., Lists 1 and 2 contain the same 50 final words, however *SPOON* is presented in an 'HC' sentence in List 1 and in an 'LC' sentence in List 2). The stimuli were generated to allow for examination of the effects of context on the processing of a final target word.

The R-SPiN corpus is a revision of the original Speech Perception in Noise corpus (SPiN; Kalikow et al., 1977), which was developed as a tool to assess the contributions of bottom-up versus top-down processes involved in understanding speech (Bilger et al., 1984). The predictability of 'HC' keywords was originally tested by presenting two groups of 12 participants with truncated written sentences (i.e.,

missing the final word), and instructing them to “fill in the word that you think is most likely to occur at the end”. Participants were informed that all responses should be monosyllabic. Sentences were removed either if all participants responded with the intended target (being deemed 'too predictable'), or fewer than two participants in either subgroup responded with the intended target (being deemed 'not sufficiently predictable'). Hence ‘HC’ stimuli included cloze values of approximately 17-92% (‘LC’ stimuli were not tested for predictability). The remaining 500 sentences were then divided across 10 lists (each including 25 ‘HC’ and 25 ‘LC’ sentences), and the intelligibility of each stimulus assessed via a speech-in-noise test of audiometrically normal hearing adults to ensure lists were equivalent (Kalikow et al., 1977). Bilger et al (1984) extended this work for adults with hearing loss. They did not retest predictability but did reassess intelligibility, leading to the removal of 50 additional target words (i.e., 100 sentences), and redistribution into lists based on this data from 128 listeners. The new lists were then validated with 32 of these listeners. The mean performance (ability to repeat the final word of each sentence) at +8 dB SNR was 76% for the ‘HC’ and 37% for the ‘LC’ sentences.

The R-SPiN procedure has been modified and updated in numerous ways since its development. For example, while originally the R-SPiN test involved presenting lists at a single SNR, typically +8 dB, Pichora-Fuller et al (1995) presented the R-SPiN using an adaptive SNR paradigm in combination with a working memory task to investigate the interaction between working memory capacity and speech in noise perception. The R-SPiN has also been reconfigured into a multiple SNR paradigm to determine the 50% recognition threshold for ‘HC’ and ‘LC’ sentences for use with listeners with audiometrically normal hearing and hearing loss (Wilson et al., 2012). Furthermore, to address potential issues of generalisability and for use in the UK, R-SPiN stimuli have been rerecorded in British English using both male and female talkers (Ward et al., 2019).

Prior findings with R-SPiN

Since its development, the R-SPiN has been used ubiquitously to understand how people perceive speech in adverse listening situations. This work has examined a range of processes including use of semantic prediction (Dubno et al., 2000; Patro & Mendel, 2016), spatial release from masking (Avivi-Reich et al., 2014), and modulation detection (Humes et al., 2013). Importantly, the R-SPiN has also been used to examine differences between top-down processing in younger and older listeners. Some research suggests that although older adults typically have more difficulty understanding speech in noise, they derive more benefit from predictive content when listening to speech in noise (Pichora-Fuller et al., 1995), noise-vocoded sentences (Sheldon et al., 2008) and sentences distorted by temporal jittering (Pichora-Fuller et al., 2007). Contradictorily, other studies find that despite differences in overall performance, older and younger listeners derive equal benefit from predictive content for interrupted speech (Kidd & Humes, 2012), and for speech in noise when audibility is controlled (Dubno et al., 2000). Hence although this is a productive area of research, age-related differences in the benefit gained from linguistic predictability remain unclear.

Shortcomings and need for re-examination

While the R-SPiN test is a useful tool for research examining the involvement of bottom-up and top-down processes in speech processing (Pichora-Fuller, 2008; Dubno et al., 2000), and it has been widely used in the field of hearing research, after 40 years, it is necessary to re-examine the R-SPiN corpus to ensure its continued relevance and validity. In such a re-examination, several potential limitations are evident, arising both from general language phenomena, and from the particulars of the original derivation of the R-SPiN stimuli by Kalikow et al (1977) and Bilger et al (1984).

Firstly, the stimuli were developed with final target words selected from the 30,000 most frequently used words prior to 1952 (Kalikow et al., 1977). Although the words may still be familiar, changes in language use over the decades could have influenced the frequency with which some words are used

(Lahar et al., 2004). As a partial but sufficient demonstration of such a trend, we conducted an exploratory analysis of change in word frequencies for those of the R-SPiN target words that are reported in the Corpus of Contemporary American English (COCA; Davies, 2008-). We found significant changes in word usage between 1990 and 2019 (dates selected as the earliest and most recent years included in the COCA). A summary of changes in word frequencies is presented in Table 1. The frequencies of target words per million were lower in 2019 compared to 1990; $Z = -2.632$, $p < .01$, meaning that many of the target words in the original corpus are no longer in such common usage, and therefore that those targets may no longer be the most appropriate for the current population.

Strikingly, use of the R-SPiN sentences today continues to rely on cloze testing completed over four decades ago, which may impact the probability of current participants selecting the 'correct' response.

Secondly, the original sentence stimuli were developed with American English listeners. Cultural differences in language may influence the predictability of some target words (e.g., DIME), thus compromising R-SPiN's validity for non-American populations (Arcuri et al., 2001).

Thirdly, whilst the R-SPiN stimuli were revised by Bilger et al. using a wide age range of listeners with hearing loss, the original sentences were developed by Kalikow et al.with input specifically from younger participants (mean age 16 years in one subtest and 17 in another). Several studies examining cloze probability and sentence constraint (Lahar et al., 2004; Arcuri et al., 2001) have raised concerns about cohort-related differences in predictions. While some such studies have demonstrated consistency between adult age groups in cloze ratings (Federmeier et al., 2002; Häuser et al., 2019), others report greater similarity in probabilities of dominant response for groups closer in age (Lahar et al, 2004). These sentences may therefore not be equivalently constraining for older adults, potentially introducing a confound when investigating age effects.

Fourthly, sentences for which at least 2 of 12 participants (17%) responded with the same key word were classed as 'High Context' and retained as such (Kalikow et al., 1977). No justification was given for

this choice of threshold, but the resulting grouping together of stimuli that could be correctly guessed almost 100% of the time, with those guessable as little as 20% of the time, could be a source of distortion or misinterpretation of results, depending on the use to which the R-SPiN corpus is being put. More recently, Block & Baldwin (2010) proposed 0-37% as an appropriate threshold for low cloze and 67-100% as an appropriate threshold for high cloze probability, as it reflects a high level of constraint (and this was validated using an event-related potential paradigm). Whilst other criteria for high cloze have been suggested, the criterion established by Block & Baldwin has subsequently been used in a number of studies (Winn & Teece, 2021; Rossi et al., 2020; Valdés Kroff et al., 2020; Luke & Christianson, 2016).

Finally, originally predictability of the final word was assessed by having participants respond (to a written sentence with final word removed) with the 'most likely' *monosyllabic noun* to complete the sentence. This stipulation may have prevented highly predictable multi-syllable words from being provided, generating spuriously high levels of agreement between respondents due to the monosyllable constraint. Compared to the target words thus categorized as 'HC', listeners might have quite different (and mutually divergent) expectations when listening to the sentences naturally. Seen another way, R-SPiN responses are only valid if the monosyllable final word constraint is fully adhered to.

Aside from general issues of language drift over time, the original R-SPiN corpus is therefore potentially burdened with the following three threats to validity:

- Incomplete transferability of corpus items across English-language cultures
- Compromised equivalence of predictability for younger and older listeners
- Low cloze threshold for 'High Context'

As noted by (Ward et al., 2019) the acoustical rendition of the corpus is an additional potential threat to validity (e.g., talker accent and gender etc.), but we focus on the linguistic content rather than issues relating to their recordings in this paper.

Aims of this study

To address the appropriateness of any corpus, it is important that the stimuli are regularly updated and assessed for the population with which it is being used. In this study, we presented R-SPiN sentences in written form, with final word removed to both young and old adults in the UK and the US, to assess the continued appropriateness of Kalikow et al (1977)'s final target word classifications for each of these four population sub-groups. From the data obtained, we derive new sentence completion norms for each of our sub-groups. For future research, and according to purpose, these can be used to compile revised lists of sentences forming reliably distinct high- and low-predictability (HP and LP) pairs, with known levels of equivalence amongst the population sub-groups studied here (UK vs. US English cultures and younger vs. older adults).

Methods

Participants

Criteria for taking part were English as a first language, age of over 60 for the older group, and age between 18-30 for the younger group, with nationality being either the UK or the US. The UK cohort included 60 (43 female) older adults (UK-O; Mean Age = 67.7, SD = 4.66) and 60 (44 female) younger adults (UK-Y; Mean Age = 25.2, SD = 3.22). The US cohort included 60 (38 female) older adults (US-O; Mean Age = 68.3, SD = 4.75) and 60 (37 female) younger adults (US-Y; Mean Age = 25.1, SD = 3.75). Ethics approval was obtained from the University of Nottingham Faculty of Medicine and Health Science Research Ethics Committee (REC reference: FMHS 423-1221). Participants provided written informed consent to participate in this study.

Stimuli

Stimuli were text renditions of the 200 'HC' and 200 'LC' sentence frames from the Revised-Speech in Noise corpus (R-SPIN; Bilger et al., 1984). Sentence frames refer to stimuli presented with the final

target word removed. We divided the stimuli into two surveys (each including 50% of the stimuli), as the R-SPiN corpus includes each final keyword occurring in both a 'HC' and 'LC' form of context. To avoid repetition of potential keywords within a survey (and thus for an individual participant), each survey therefore only included one version of each R-SPiN pair (i.e., for SPOON, *He stirs the coffee with a* ____ was presented in Survey 1, whereas *Bob could have known about the* ____ was presented in Survey 2).

Procedure

Participants were recruited via Prolific (www.prolific.com) and the task was presented online via JISC (JISC, 2020). Stimuli were shown as written sentences with the final target word replaced by a blank line. Participants were asked to "complete each sentence by typing only one word", i.e., a Cloze task. Participants responded to 200 sentences, and no time limit was imposed. Participants were required to provide a response for every sentence, but not specifically told that they had to provide a monosyllabic response.

Note that as each participant completed the Cloze task for 50% of the original R-SPiN items. (As each keyword was used in both 'HC' and 'LC' sentences, stimuli were divided so participants only encountered each keyword once, thus seeing 100/200 'HC' and 100/200 'LC' sentence frames each.) This led to 30 participants per sub-group providing a Cloze task response for each stimulus.

Analysis

All responses were first checked and corrected for errors in spelling. Differences in tense, plurality or suffixes were counted as the same word (e.g., clap and claps, or clap and clapping), as were words with minor misspellings (e.g., CLOC rather than CLOCK).

However, misspelled words where it wasn't clear what the intended word was (e.g., for the sentence frame *Mary wore her hair in* ___, the given response 'Hait' could have been intended as *PLAITS* but is different enough to be ambiguous), were excluded from analysis, resulting in a loss of 0.05% responses

from the UK old group, 0.03% from the UK young group, and 0.02% from the US young group. Differences in capitalization (e.g., *CLOCK* or *clock*), and punctuation (e.g., *CLOCK?*) were ignored. American and English variations on spelling (e.g., *JEWELRY* or *JEWELLERY*, *MOLD* or *MOULD*) were counted as the same word. Where the participant responded with more than one word as an answer (e.g., *FIRST MATE*, *COFFEE TABLE*), that response was excluded from analysis, resulting in a loss of 0.57% data for the UK older group, 0.19% for the UK young group, 0.08% for the US older group, and 0.17% for the US young group (0.25% of the data in total). (Note that one participant in the UK older adult group was replaced as they responded with > 1 word for more than 25% of stimuli, and one participant was replaced in the US young adult group as they responded with the same word for 64% of stimuli.)

Response probability statistics were calculated separately for each of the four sub-groups. Cloze probability for a given sentence frame plus response was calculated as the number of participants giving that response divided by the number of participants giving valid responses for the given sentence frame. In line with Block & Baldwin (2010) and Winn & Teece (2021), we define high cloze probability as at least 67%, and low cloze probability as up to 33%. We will refer to sentences as highly predictable (HP) if responses met the 67% criterion for high cloze, and as having low predictability (LP) if the responses met the 33% criterion for low cloze. Note that the 67% threshold guarantees that only one response word can be classified as HP for a given sentence frame.

In addition, the number and probability of all unique responses to each sentence frame was used to assess response entropy (Shannon & Weaver, 1949). Entropy is a function of the distribution of probabilities of all possible responses ($- \Sigma\, p(x)*\log_2 p(x)$), where x is probability p of each unique response occurring). Entropy is low when one response is more probable than others (i.e., participants converge on the same word) and high when multiple responses are equally probable. As such, higher entropy is associated with more uncertainty about possible sentence final word completions. Different levels of entropy may influence how words with similar levels of cloze probability are processed.

Predictability and response entropy are compared between the four participant groups to reveal any differences associated with age and nationality. We first identify stimuli that pass our cloze threshold of 67% in each group. T-tests are used to compare cloze values and response entropy between groups. Subsequently, we derive subsets of matched sentence pairs (satisfying the above criteria for HP and LP). These subsets will differ according to whether they are to be consistent across age groups, nationalities, or both, and whether or not they are restricted to sentence-final words present in the original R-SPiN corpus.

Results

An example of observed responses for an 'HC' and its matched 'LC' R-SPiN sentence frame, including cloze probabilities for each participant sub-group, is shown in Table 2. Corresponding data for all stimuli are presented in Supplementary Material.

Cloze Probability

UK Older Adults: A total of 109 of the 200 'HC' stimuli met the criteria for high cloze probability (ranging from 70 – 100%). Of these 109 HP sentence frames, 96 responses matched the target word in the original R-SPiN stimuli. Of 200 'LC' stimuli, four did not meet the < 33% threshold for LP.

UK Younger Adults: A total of 75 of the 200 'HC' stimuli met the criteria for high cloze probability (ranging from 70 – 100%). Of these 75 HP sentence frames, 62 responses matched the target word in the original R-SPiN stimuli. Of 200 'LC' stimuli, two did not meet the < 33% threshold for LP.

US Older Adults: A total of 107 of the 200 'HC' stimuli met the criteria for high cloze probability (ranging from 70 – 100%). Of these 107 HP sentence frames, 99 responses matched the target word in the original R-SPiN stimuli. Of 200 'LC' stimuli, two did not meet the < 33% threshold for LP.

US Younger Adults: A total of 81 of the 200 'HC' stimuli met the criteria for high cloze probability (ranging from 70 – 100%). For two stimuli meeting our high cloze threshold their 'LC' pair did not meet

the < 33% threshold for LP. Of the 79 remaining HP sentence frames, 67 responses matched the target word in the original R-SPiN stimuli. Of 200 'LC' stimuli, only the previously mentioned two stimuli did not meet the < 33% threshold for LP.

Overall: Only 48 of the 200 'HC' stimuli scored over 67% in our cloze task for all four participant groups. Of these, 45 stimuli were completed with the same final word across groups, and 44 were completed using the final word of the original R-SPiN corpus. These 44 words are listed in Table 3.

With age and culture groupings not taken into account and data collapsed across all 120 participants, a total of 86 of the 200 'HC' stimuli met the criteria for HP (ranging from 67.23 – 98.33%), 80 of which were completed with the final word in the original R-SPiN.

Language background effects

Cloze probability for 'HC' stimuli was calculated separately for all participants in the UK group and all in the US group. Cloze values for 200 'HC' stimuli did not significantly differ between UK (M = 62.27%, SD = 22.14%) and US (M = 63.58%, SD = 20.92%) groups; $p = .227$, although dominant responses varied for 37 of the 200 stimuli. Dominant responses in the US group matched more target words in the R-SPiN (159) than for the UK group (138). The number of stimuli that passed our 67% threshold for high cloze probability were 87 in the UK group and 88 in the US group. Of those, 75 for the UK group and 83 for the US were completed using the final word of the original R-SPiN corpus. Cloze values of responses that matched the original target word in R-SPiN (ranging from 15 – 100%) were significantly higher in the US (M = 56.88%, SD = 27.65%) than the UK (M =51.97%, SD = 30.71%) groups; $t(199) = -4.151$, $p < .001$.

Age effects

For the UK groups, the average cloze value of 'HC' stimuli was significantly higher for older (M = 67.75%, SD = 22.55 %) versus younger (M = 59.52%, SD = 23.25 %) participants; $t(199) = 6.141$, $p < .001$. There were 62 stimuli that passed our high cloze threshold for both older and younger participants, with both

groups agreeing the same final word. Of these 62 HP words, 56 matched the final word in the original R-SPIN stimuli.

For the US groups, the average cloze value of 'HC' stimuli was significantly higher for older (M = 69.53%, SD = 21.22%) versus younger (M = 59.41%, SD = 22.27%) participants; $t(199) = 8.370$, $p < .001$. There were 67 out of 200 stimuli that passed our high cloze threshold in both age groups, although the word with the highest proportion of responses differed across groups for one stimulus. Of the remaining 66 words that were consistent across US groups, 62 matched the final word in the R-SPiN stimuli.

Response entropy

The entropy values for 'HC', 'LC', and HP sentence stimuli for each group are summarized in Table 4. Note that lower entropy scores result from two factors: fewer unique responses to a stimulus being provided and more variability in the distribution of their probabilities, and therefore higher entropy indicates less agreement between participants.

Language background effects

Response entropy for 'HC' stimuli was not significantly lower for the US (M=1.76, SD=.91) versus UK (M=1.78, SD=.95) group; $p = .650$. Entropy for 70 'HC' stimuli that passed the 'HP' threshold in both groups was not significantly different between US (M=0.90, SD=0.43) compared to the UK (M=0.87, SD=0.48) group; $p = .452$. However, entropy for 'LC' stimuli was significantly lower in the US (M=5.31, SD=0.32) than the UK (M=5.20, SD=0.36) group; $t(199)= -.756$, $p < .001$, indicating greater agreement between US participants for these stimuli than UK participants.

Age effects

For the UK groups, response entropy for 'HC' sentences was significantly lower in Older (M = 1.38, SD = 0.87) versus Younger (M = 1.78, SD = 0.98) adults; $t(199) = -7.231$, $p < .001$, indicating greater agreement between older participants. Entropy for the 62 stimuli that passed the 'HP' threshold for both age groups was also significantly lower for Older (M = 0.57, SD = 0.39) than the Younger (M = 0.73, SD =

0.48) adults; t(61) = -2.377, p < .05. Entropy did not significantly differ between age groups for 'LC' sentences.

For the US groups, response entropy for 'HC' sentences was again significantly lower for Older US (M = 1.33, SD = 0.86) versus Younger US (M = 1.83, SD = 0.93) groups; t(199) = -9.800, p < .001. Entropy for 67 sentences that passed the high cloze threshold in both US age groups was also significantly lower for Older (M = 0.54, SD = 0.38) than for Younger (M = 0.86, SD = 0.35) adults; t(66) = -6.107, p < .001. Response entropy for 'LC' sentences did not significantly differ between age groups.

Revised stimulus sets

Based on the complete dataset, we can derive lists of matched sentence pairs satisfying the above criteria for both HP and LP, suitable for use across nationalities (UK vs. US), age groups, or both. Here we present each list in two versions; a short version only allowing sentence-final words found in the original R-SPiN corpus, and a longer version additionally allowing for sentences to be completed with other words (e.g., allowing POUND or DOLLAR instead of the original final word DIME for the sentence How much can I buy for a ____). The ten lists are given in full in Supplementary Material. Table 5 summarises their characteristics.

In developing the original R-SPiN stimuli, Kalikow et al (1977) excluded from their 'HC' list any words showing 100% cloze probability, on the basis that the final word would always be identifiable through contextual repair, regardless of masking of that final word. An argument can be made that 100% is not uniquely inappropriate, as stimuli with cloze probabilities 'close to 100%' will also distort results in practical use. Furthermore, while 100% cloze is uninformative if one is studying effects of predictive context on identification of a degraded final word, if the study concerns the effects of degrading the predictive context itself (or the whole sentence), then including words with 100% cloze is acceptable. For these reasons, we have refrained from excluding words with cloze probabilities above any specific

371 threshold. The user of our revised lists is encouraged to apply them according to purpose, and consult
372 Supplementary Material to determine lists appropriate to their use as relevant.

373 Discussion

374 The aim of the study was to assess whether R-SPiN stimuli are still appropriate for a contemporary
375 population, across two English-language cultures, and both younger and older adults. To address this,
376 we presented the R-SPiN stimuli (in text form) with final word removed to older and younger adults in
377 the UK and the US and asked them to complete each sentence frame.
378 Our findings show that many of the 200 original R-SPiN stimuli designated as 'High context' do not meet
379 the criteria of HP (i.e., cloze >= 67%). Between the four participant groups the number of stimuli classed
380 as HP ranged from 75 to 108 (out of 200). Strikingly, only 48 met the HP criteria across all groups, and
381 only 45 when requiring the same predicted final word.
382 Although average cloze probability, collapsed across age groups, for all 200 'HC' stimuli was not
383 significantly different between the UK and US groups (UK: 62.32%; US: 63.50%), dominant responses
384 varied for a number of sentences. Cloze probabilities of all responses matching the target word in the R-
385 SPiN corpus were greater in the US than the UK group (UK: 52.03%; US: 56.80%). Only entropy for low
386 context sentences was significantly different between UK and US. Hence caution is also warranted when
387 using the same word sets across language backgrounds (even disregarding any effects of accent).
388 Strikingly, we observed an age effect (or perhaps more correctly, a birth cohort effect) for both cloze
389 probability and entropy. In terms of cloze probability, more of the 'HC' stimuli were predicted by older
390 adults (UK: 108; US: 106) versus younger adults (UK: 75; US: 80); in terms of entropy, 'HC' sentences
391 elicited fewer responses on average for older (UK: 5.51; US: 5.39) versus younger adults (UK: 7.10; US:
392 7.56). The issue of stimuli not being equivalently constraining (i.e., being more predictable for older than

younger adults) may therefore be a contributing factor for inconsistent age effects reported in the literature.

Following these analyses, we also want to highlight that a given value of predictability (i.e., cloze probability) does not uniquely determine how easily or quickly a sentence will be processed. In the present study, both *The crook entered a guilty plea* and *To open the jar, twist the lid* elicited cloze values of 85.83% (averaged across groups), yet there was greater response entropy for the former (14 unique responses versus 3, respectively). Different levels of constraint on possible continuations could lead to words with equivalent levels of predictability being processed differently.

In the original selection of keyword characteristics for SPiN sentences, it was decided that the final target word should always be a monosyllabic noun. This requirement was justified by Kalikow et al (1977) as necessary in order to maintain a degree of control over phonetic and prosodic factors. However, in their subsequent testing of key word predictability, participants were informed that the target words were all monosyllabic nouns. This may have led to some sentence stimuli being misleadingly categorized as 'High Context', due to the sentence constraint allowing only a single monosyllabic completion, though potentially many multi-syllabic completions could be provided. For example, 'couch' vs 'sofa'/'armchair'. In our testing of the stimuli, participants were not given any stipulations as to the number of syllables the final word should contain. On average 22% of all sentence completions for 'HC' stimuli were multi-syllabic (mainly the non-dominant completions), ranging from 0-96% across individual sentence frames. Only 10% of all responses to stimuli that passed our 67% cloze threshold for High Predictability were multi-syllabic, compared to 30% multi-syllabic responses to stimuli that did not pass the threshold. For stimuli that met the HP threshold, multi-syllabic words made up less than 5% of dominant responses in any group (UK-Old: 4.59%; UK-Young: 1.33%; US-Old: 3.74%; US-Young: 6.17%). Although stimuli may then have different numbers of syllables in their final word, the inclusion of multi-syllabic responses provides a better estimate of response uncertainty (entropy) when

listening naturally, and therefore better reflects the likelihood of any one word being predicted over others (even if the dominant response ends up being monosyllabic). For example, while *tea* was the original R-SPiN target for *Ruth poured herself a cup of* ___, when allowing multi-syllable responses in our cloze test we found older US participants to be almost evenly split between *tea* and *coffee* (while UK participants were consistent in their choice of *tea*).

The revised lists facilitate the generation of reliable stimuli sets to investigate age-related effects in the speech processing of older and younger adults in the UK and US for at least a few years hence. This provide a basis for future research pursuing the resolution of existing conflicting results.

Limitations

It is important to note that generating HP lists from our tests has some limitations. In terms of experimental considerations, not all HP final words have the same number of syllables (ranging from 1-5), which may be inappropriate for use in tests of speech intelligibility. In addition, applying a 67% cloze threshold greatly reduces the number of usable stimuli, with the appropriate subset varying with the intended population. Indeed the subset found to be reliable across both UK and US samples for young and old adults only contains 44 items, against the 200 in the original R-SPiN corpus. Finally, although we have attempted to update the stimuli with more appropriate final words, in some cases the sentence frame itself is outdated either due to certain words falling out of use (e.g., *crook*) or to general changes in technology and behaviour (e.g., *My TV has a 12-inch screen*). We note furthermore that if our amended lists were used it would be important to re-record them with appropriate final words, and to assess the intelligibility of the stimuli with the new final words. Should the stimuli be broken down into separate lists, these should also be tested for equivalence.

We also note that only included participants between 19-31 for the younger group, and 60-80 for the older group. Based on the findings reported here and by previous studies evaluating cloze differences (albeit for different corpora) in other populations (e.g., in children and adolescents, Rossi et al., 2020;

Pinheiro et al., 2010, and in young, middle-aged, and older adults, Lahar et al., 2004), we expect predictions may differ for listeners falling outside of these age groups. We therefore want to further emphasise the importance of cloze testing stimuli in the population being studied.

Conclusions

The present study aimed to examine the suitability of the R-SPiN stimuli for use amongst different age groups and English-language cultures in the present day. We were particularly concerned about three potential threats to current-use validity: issues of language culture, age, and high context threshold. Although our UK and US groups showed a similar overall cloze for 'High Context' sentence frames, we noted variation between dominant word choice both between groups and in relation to the original R-SPiN stimuli, highlighting the issue of incomplete transferability across English-language cultures. Furthermore, we found compromised equivalency of predictable items across age, with older adults providing more consistent responses to R-SPiN's so-called 'High Context' stimuli. Finally, we found that fewer than half the stimuli elicited a 67% agreement on the most likely continuation, and the subset found to be reliable across English-language cultures and age only contained 44 items. Overall, these findings emphasize the potential influences of age, culture and time of testing on expected continuations of R-SPiN sentences.

We provide lists of subsets of stimuli from the R-SPiN corpus whose items show at least 67% cloze probability (i.e., High Predictability) across age groups (young vs. old), national groups (UK vs. US), or both, as Supplementary Materials. We also provide cloze probabilities and entropy information for each stimulus. We hope this data will spark further exploration of age and language-culture effects of context, including further exploration of previously collected data, and robust future work.

Data availability Statement

The datasets generated and analysed during the current study are available from the corresponding author on request. All lists generated are included in the supplemental files.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Arcuri, S. M., Rabe-Hesketh, S., Morris, R. G., & McGuire, P. K. (2001). Regional variation of cloze probabilities for sentence contexts. *Behavior Research Methods, Instruments, & Computers*, *33*, 80–90.

Avivi-Reich, M., Daneman, M., & Schneider, B. A. (2014). How age and linguistic competence alter the interplay of perceptual and cognitive factors when listening to conversations in a noisy environment. *Frontiers in Systems Neuroscience*, *8*. https://www.frontiersin.org/articles/10.3389/fnsys.2014.00021

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech, Language, and Hearing Research*, *27*(1), 32–48.

Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, *42*(3), 665–670. https://doi.org/10.3758/BRM.42.3.665

Bloom, P. A., & Fischler, I. (1980). Completion norms for 329 sentence contexts. *Memory & Cognition*, *8*(6), 631–642. https://doi.org/10.3758/BF03213783

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–2538.

Davies, M. (2008). *The corpus of contemporary American English: 450 million words, 1990-present.*

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

Drijvers, L., & Özyürek, A. (2017). Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Dubno, J. R., Ahlstrom, J. B., & Horwitz, A. R. (2000). Use of context by young and aged adults with normal hearing. *The Journal of the Acoustical Society of America*, *107*(1), 538–546.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., McLennan, D. B., Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, *39*(2), 133–146. https://doi.org/10.1111/1469-8986.3920133

Golestani, N., Rosen, S., & Scott, S. K. (2009). Native-language benefit for understanding speech-in-noise: The contribution of semantics. *Bilingualism: Language and Cognition*, *12*(3), 385–392.

Häuser, K. I., Demberg, V., & Kray, J. (2019). Effects of Aging and Dual-Task Demands on the Comprehension of Less Expected Sentence Continuations: Evidence From Pupillometry. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00709

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639–652.

Humes, L. E., Kidd, G. R., & Lentz, J. J. (2013). Auditory and cognitive factors underlying individual differences in aided speech-understanding among older adults. *Frontiers in Systems Neuroscience*, *7*, 55.

JISC. (2020). *JISC Online Surveys [Software]*.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995–2004.

Johnsrude, I. S., & Rodd, J. M. (2016). Factors That Increase Processing Demands When Listening to Speech. In *Neurobiology of Language,* 491–502. Elsevier. https://doi.org/10.1016/B978-0-12-407794-2.00040-7

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337–1351.

Keil, J., Müller, N., Ihssen, N., & Weisz, N. (2012). On the variability of the McGurk effect: audiovisual integration depends on prestimulus brain states. *Cerebral Cortex*, *22*(1), 221–231.

Kidd, G. R., & Humes, L. E. (2012). Effects of age and hearing loss on the recognition of interrupted words in isolation and in sentences. *The Journal of the Acoustical Society of America*, *131*(2), 1434–1448.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kutas, M., & Federmeier, K. D. (2010). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Lahar, C. J., Tun, P. A., & Wingfield, A. (2004). Sentence–final word completion norms for young, middle-aged, and older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *59*(1), 7–10.

Lash, A., Rogers, C. S., Zoller, A., & Wingfield, A. (2013). Expectation and Entropy in Spoken Word Recognition: Effects of Age and Hearing Acuity. *Experimental Aging Research*, *39*(3), 235–253. https://doi.org/10.1080/0361073X.2013.779175

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: behavioural and neural interactions. *Neuropsychologia*, *50*(5), 762–776.

Obermeier, C., Dolk, T., & Gunter, T. C. (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex*, *48*(7), 857–870. https://doi.org/10.1016/j.cortex.2011.02.007

Obleser, J. (2014). Putting the Listening Brain in Context. *Language and Linguistics Compass*, *8*(12), 646–658. https://doi.org/10.1111/lnc3.12098

Obleser, J., & Kotz, S. A. (2010). Expectancy constraints in degraded speech modulate the language comprehension network. *Cerebral Cortex*, *20*(3), 633–640.

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *Neuroimage*, *55*(2), 713–723.

Obleser, J., Wise, R. J. S., Dresner, M. A., & Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience*, *27*(9), 2283–2289.

Patro, C., & Mendel, L. L. (2016). Role of contextual cues on the perception of spectrally reduced interrupted speech. *The Journal of the Acoustical Society of America*, *140*(2), 1336–1345.

Pichora-Fuller, M. K. (2008). Use of supportive context by younger and older adult listeners: Balancing bottom-up and top-down information processing. *International Journal of Audiology*, *47*:sup2, S72–S82. https://doi.org/10.1080/14992020802307404

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608.

Pichora-Fuller, M. K., Schneider, B. A., MacDonald, E., Pass, H. E., & Brown, S. (2007). Temporal jitter disrupts speech intelligibility: A simulation of auditory aging. *Hearing Research*, *223*(1–2), 114–121.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044. https://doi.org/10.1037/bul0000158

Pinheiro, A. P., Soares, A. P., Comesaña, M., Niznikiewicz, M., & Gonçalves, Ó. F. (2010). Sentence-final word completion norms for European Portuguese children and adolescents. *Behavior Research Methods*, *42*(4), 1022–1029.

Rossi, N. F., Fernandes, C., Moreira, C. S., Giacheti, C. M., Sichieri, B. B., Pinheiro, A. P., & Sampaio, A. (2020). Sentence contexts and cloze probabilities for Brazilian Portuguese children and adolescents. *PLOS ONE*, *15*(7), e0236388. https://doi.org/10.1371/journal.pone.0236388

Schneider, B. A. (2011). How Age Affects Auditory-Cognitive Interactions in Speech Comprehension. *Audiology Research*, *1*(1), e10. https://doi.org/10.4081/audiores.2011.e10

Schneider, B. A., and Pichora-Fuller, M. K., and Daneman, M. (2010). Effects of Senescent Changes in Audition and Cognition on Spoken Language Comprehension. *The Aging Auditory System*, 167–210. Springer New York. https://doi.org/10.1007/978-1-4419-0993-0_7

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication. *University of Illinois Press*.

Sheldon, S., Pichora-Fuller, M. K., & Schneider, B. A. (2008). Priming and sentence context support listening to noise-vocoded speech by younger and older adults. *The Journal of the Acoustical Society of America*, *123*(1), 489–499.

Valdés Kroff, J. R., Román, P., & Dussias, P. E. (2020). Are All Code-Switches Processed Alike? Examining Semantic v. Language Unexpectancy. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.02138

van der Feest, S. V. H., Blanco, C. P., & Smiljanic, R. (2019). Influence of speaking style adaptations and semantic context on the time course of word recognition in quiet and in noise. *Journal of Phonetics*, *73*, 158–177. https://doi.org/10.1016/j.wocn.2019.01.003

Van Engen, K. J., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2014). Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1908–1918.

Ward, L., Robinson, C., Paradis, M., Tucker, K. M., & Shirley, B. G. (2019). R2SPIN: Re-Recording the Revised Speech Perception in Noise Test. *INTERSPEECH*, 3133–3137.

Wilson, R. H., McArdle, R., Watts, K. L., & Smith, S. L. (2012). The Revised Speech Perception in Noise Test (R-SPIN) in a multiple signal-to-noise ratio paradigm. *Journal of the American Academy of Audiology*, *23*(08), 590–605.

Winn, M. B. (2016). Rapid Release From Listening Effort Resulting From Semantic Context, and Effects of Spectral Degradation and Cochlear Implants. *Trends in Hearing*, *20*. https://doi.org/10.1177/2331216516669723

Winn, M. B., & Teece, K. H. (2021). Listening Effort Is Not the Same as Speech Intelligibility Score. *Trends in Hearing*, *25*. https://doi.org/10.1177/23312165211027688

Table 1. Summary of target words in the original R-SPiN corpus that differed in frequency in the COCA between the years 1990 and 2019 (Positive differences indicate greater use in 2019 versus 1990. Negative differences indicate less frequent use in 2019 versus 1990. Ties indicate no difference in word frequency).

	N
Negative Differences	123
Positive Differences	76
Ties	1
Total	200

TABLE 2. Excerpt of all responses for 'HC' and 'LC' sentences having R-SPiN target word DIME, with observed cloze probabilities for each sub-group.

Sentence Frame	R-SPiN Target Word	Unique Responses	Cloze Probability (proportion of participants providing each response)				
			UK		US		
			Old	Young	Old	Young	All
'High Context'							
How much can I buy for a	**DIME**						
		DOLLAR		3.33%	86.67%	83.33%	43.33%
		POUND	83.33%	73.33%			39.27%
		FIVER	10.00%	10.00%			5.00%
		TENNER		6.67%			1.67%
		QUARTER			3.33%	3.33%	1.67%
		DIME			6.67%		1.67%
		NICKEL			3.33%	3.33%	1.67%
		Other (7)	6.67%	6.67%	0.00%	10.00%	5.83%
		N responses	4	6	4	6	14
'Low Context'							
You want to think about the	**DIME**						
		FUTURE	6.67%	13.33%	6.67%	16.67%	10.83%
		CONSEQUENCES	6.67%	13.33%	3.33%	3.33%	6.67%
		PROBLEM	3.33%		6.67%	3.33%	3.33%
		GAME				13.33%	3.33%
		CHILDREN	6.67%		6.67%		3.33%
		IDEA	3.33%	3.33%		3.33%	2.50%
		JOB		3.33%	6.67%		2.50%
		DECISION		3.33%		6.67%	2.50%
		OUTCOME	3.33%	3.33%			1.67%
		OTHERS	3.33%			3.33%	1.67%
		OPTIONS		3.33%	3.33%		1.67%
		PLAN		3.33%		3.33%	1.67%
		DAY		3.33%		3.33%	1.67%
		MEAL		3.33%		3.33%	1.67%
		COST	6.67%				1.67%
		RISK	6.67%				1.67%
		WEATHER		6.67%			1.67%
		DOG		6.67%			1.67%
		GIRL				6.67%	1.67%
		Other (56)	53.33%	33.33%	66.67%	33.33%	46.67%
		N responses	25	22	26	21	75
N.B. Responses provided by only one participant across all groups are collapsed under Other							

N=30 participants, hence cloze probability 3.33% represents one response.

TABLE 3. Shows the 44 target words that were classified as HP for each of our population groups.

Spoon	Coin	Calf	Lanes	Clock	Sword	Map	Screen	Crown	Frogs	Fist
Wheels	Trap	Flood	Wrist	Rent	Vest	Belt	Mouse	Plea	Fur	Lid
Jar	Mice	Mold	Breath	Sleeves	Hay	Pole	Pork	Throat	Roar	Stripes
Slice	Wax	Fans	Sand	Shell	Knife	Row	Sheep	Thorns	Track	Blade

TABLE 4: Average cloze probability and response entropy (number of unique responses) for 'High Context', 'Low Context', High Predictability and Low Predictability sentences

	'High Context' Stimuli ('HC')			'Low Context' Stimuli ('LC')			High Predictability Stimuli (HP)			Low Predictability Stimuli (LP)		
Group	**Avg. Cloze**	**N Unique Responses**	**Entropy**	**Avg. Cloze**	**N Unique Responses**	**Entropy**	**Avg. Cloze**	**N Unique Responses**	**Entropy**	**Avg. Cloze**	**N Unique Responses**	**Entropy**
UK Older	67.75%	5.52	1.38	12.03%	24.30	4.43	85.67%	3.71	0.75	11.39%	24.54	4.46
UK Young	59.52%	7.07	1.78	11.62%	24.55	4.47	85.05%	3.77	0.77	11.37%	24.62	4.48
US Older	69.53%	5.38	1.33	11.23%	25.15	4.51	86.73%	3.40	0.68	10.72%	25.30	4.54
US Young	59.42%	7.53	1.83	10.00%	25.69	4.57	82.84%	4.54	0.92	9.76%	25.77	4.58

TABLE 5: Characteristics of lists derived from our data, for matched HP+LP sentences. Lists marked 'Short' include only sentence-final words from the original R-SPiN corpus, whereas lists marked 'Long' include additional sentence-final words. Lists marked 'UK-US' are usable across both nationalities, and lists marked 'Adult' are usable across both age groups.

	UK-US Adult Short	UK-US Adult Long	UK-US Old Short	UK-US Old Long	UK-US Young Short	UK-US Young Long	US-Only Short	US-Only Long	UK-Only Short	UK-Only Long
N HP words	44	45	80	82	51	57	62	66	56	62
N multi-syllabic words	0	0	0	1	0	1	0	2	0	1

Supplemental Material

Lists of matched sentence pair satisfying criteria for HP and LP suitable for use across nationalities (UK vs. US), age groups, or both. Each list exists in two versions, a short version containing only sentence-final words found in the original R-SPiN corpus, and a long version allowing for final words that do not match those in the original R-SPiN corpus.