JASA ARTICLE

Don’t force it! Gradient speech categorization calls for continuous categorization tasks[a]

Keith S. Apfelbaum,[1,b] Ethan Kutlu,[1,b] Bob McMurray,[1,b] and Efthymia C. Kapnoula[2,3,c]

[1]*Department of Psychological and Brain Sciences, G60 Psychological and Brain Sciences Building, University of Iowa, Iowa City, Iowa 52242-1407, USA*

[2]*BCBL, Basque Center on Cognition, Brain and Language, Mikeletegi 69, 20009 Donostia, Spain*

[3]*Ikerbasque, Basque Foundation for Science, Plaza Euskadi 5, 48009 Bilbao, Spain*

ABSTRACT:
Research on speech categorization and phoneme recognition has relied heavily on tasks in which participants listen to stimuli from a speech continuum and are asked to either classify each stimulus (identification) or discriminate between them (discrimination). Such tasks rest on assumptions about how perception maps onto discrete responses that have not been thoroughly investigated. Here, we identify critical challenges in the link between these tasks and theories of speech categorization. In particular, we show that patterns that have traditionally been linked to categorical perception could arise despite continuous underlying perception and that patterns that run counter to categorical perception could arise despite underlying categorical perception. We describe an alternative measure of speech perception using a visual analog scale that better differentiates between processes at play in speech categorization, and we review some recent findings that show how this task can be used to better inform our theories.
https://doi.org/10.1121/10.0015201

(Received 19 April 2022; revised 12 September 2022; accepted 20 October 2022; published online 20 December 2022)

[Editor: Richard A. Wright] Pages: 3728–3745

I. INTRODUCTION

Categorization of speech sounds requires listeners to extract linguistic information from a highly variable speech signal. The same phoneme or feature varies in its acoustic manifestation depending on talker sex, speaking rate, surrounding phonetic context, and various other factors, and the same constellation of acoustic cue values can be consistent with multiple phonemes. This *problem of lack of invariance* in the speech signal has been the focus of much research on speech perception and categorization for decades (Kluender, 1994; Liberman and Whalen, 2000; Perkell and Klatt, 2014). How do listeners translate a variable acoustic signal into more discrete units like phonological features, phonemes, or words? Given this variability, the ability to recognize speech segments has often been conceptualized as a two-stage process: first, continuous acoustic cues are encoded, and second, these cues are mapped onto categories. This means that recognizing speech segments is fundamentally a problem of categorization (Holt and Lotto, 2010)—the ability to group perceptually dissimilar tokens.

Theories of how listeners overcome variability in the continuous speech signal rely on tasks that can accurately assess how the auditory signal is encoded and later mapped to categories. However, the conceptual links between classic categorization or discrimination tasks and the underlying processes of speech categorization are often unclear, and issues with these links can lead theories astray. This is particularly the case as the field has moved from a view—grounded in categorical perception—that categorization is necessarily discrete to a more modern view that it is gradient [see also McMurray and Haskins Laboratories (2022)]. This shift raises the need for a critical reassessment of the methodologies used to understand these processes and development of methods that capture the nature of speech categorization with fidelity.

II. CATEGORICAL PERCEPTION

For many years, the dominant theory of how listeners overcome variability in the speech signal was *categorical perception* (CP) [Liberman *et al.*, 1957; Repp, 1984; for a review and critique of CP, see McMurray and Haskins Laboratories (2022)]. This theory posited that listeners rapidly discard acoustic information that is irrelevant to category identity and instead directly perceive the category itself. For example, the principal cue differentiating a /b/ from a /p/ is voice onset time (VOT): /b/s have short VOTs of ∼0 ms, whereas /p/s have longer VOTs of ∼40 ms; the boundary between the two is at around 20 ms. However, VOT varies across productions. CP suggests that despite the variability in the precise VOT of a given /b/, all tokens are perceived identically, simply as /b/. The surface variability is discarded during the initial stages of perception, such that only the category identity is encoded. While this theory does not spell out precisely how listeners cope with variability, the existence of CP as an empirical

[a]This paper is part of a special issue on Reconsidering Classic Ideas in Speech Communication.
[b]Also at: DeLTA Center, University of Iowa, 109 South Lindquist Center, Iowa City, IA 52242, USA.
[c]Electronic mail: kapnoula@gmail.com.

phenomenon suggests that listeners solve this problem by *creating the invariance themselves*.

Evidence for CP (Liberman *et al.*, 1957) came from two tasks: discrimination and forced-choice identification.

With respect to identification (Fig. 1, dotted line), CP suggests that listeners should be insensitive to within-category differences—a listener should be equally adept at identifying a token as a /b/ if it has prototypical acoustic characteristics for the /b/ category or if it falls close to the boundary with /p/. Identification should reflect largely—or in the extreme, solely—whether the stimulus falls on the /b/ or /p/ side of the boundary, no matter its VOT. This predicts sharp changes to identification as the other category for tokens on the other side of the category boundary.

Second, with respect to discrimination (Fig. 1, solid line), CP predicts that listeners only perceive category identity. Here, we use perception to specifically refer to the auditory encoding—the form that initial representations take in the earliest stages of processing. If listeners' perceptual (cue-level) encoding only reflects category identity, they should be unable to discriminate tokens within a category (e.g., two acoustically different /b/s); from the earliest stages of cognitive processing, subcategorical detail is jettisoned. At the same time, discrimination should be strong for tokens in different categories (a /b/ vs a /p/), even if the acoustic distance is held constant. This predicts a discrimination function with flat, near-chance performance for differences within categories and sharp spikes for tokens that cross the boundary (Fig. 1, solid line). This profile of discrimination is the crucial hallmark of CP.[1] Since typical discrimination tasks do not require labeling in principle, they are thought to assess the pre-categorical encoding of the auditory input or phonetic cues. Consequently, this pattern of discrimination performance is typically interpreted as evidence that the low-level perceptual encoding of speech and the mapping from encoding to categories are one and the same—categorization occurs immediately during initial stages of cognitive processing of speech stimuli.

Early work in speech categorization revealed just these patterns of discrimination and identification. This led to widespread acceptance of CP (Harnad, 1987; Repp, 1984) and its theoretical claim that listeners discard continuous acoustic detail. However, this conclusion depends on the assumption that discrimination and identification tasks directly reflect the underlying processes of categorization and auditory encoding of speech (respectively) and that, therefore, these two processes are isomorphic.

Many of the theoretical claims of CP ultimately withered in the face of later research using other methods, such as priming (Andruski *et al.*, 1994), continuous rating scales (Massaro and Cohen, 1983; Miller and Volaitis, 1989), the visual world paradigm (VWP) (McMurray *et al.*, 2002), and event-related potentials (ERPs) (Toscano and McMurray, 2010). These provided convergent evidence that listeners maintain continuous information throughout the process of speech categorization and even into word recognition. Listeners exhibit clear gradiency in their speech categorization with some tokens being categorized as more robust exemplars than others [see McMurray and Haskins Laboratories (2022) for review]. Despite this theoretical consensus, the methodological assumption about discrimination and identification tasks remains largely unquestioned; forced-choice identification and, to a lesser extent, discrimination tasks are to this day considered standard ways of measuring speech categorization and auditory encoding of speech sounds.

As we argue next, detailed analysis of how task characteristics shape these measures has long suggested a more complicated story (Carney *et al.*, 1977; Gerrits and Schouten, 2004; Pisoni and Tash, 1974; Schouten *et al.*, 2003). Both standard discrimination and forced-choice identification tasks may inadvertently incorporate other cognitive and decision processes, and these other processes might warp or mask the ability of the tasks to clearly index how listeners categorize speech sounds. This has long been discussed in the context of discrimination tasks; however,

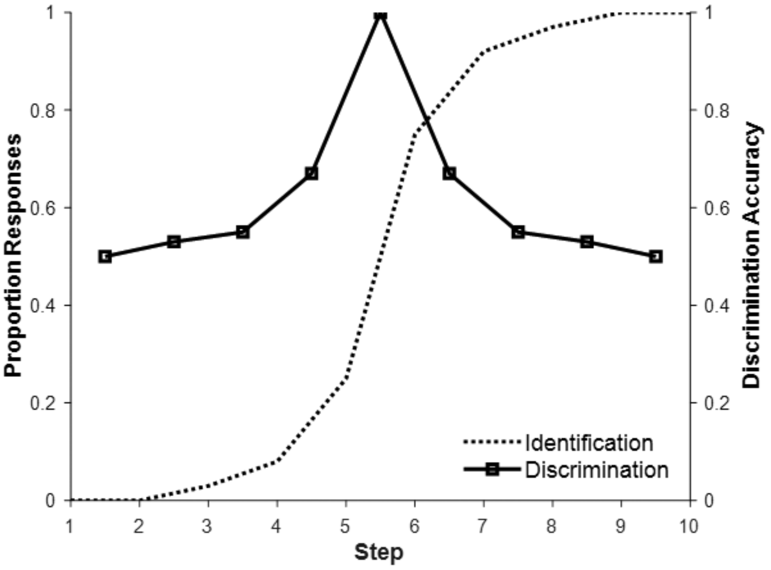


FIG. 1. Classic patterns of identification (dotted line) and discrimination (solid line) performance that led to theories of CP.

it has received less attention for identification, the focus of this review.

III. DISCRIMINATION TASKS

Substantial work has explored the underlying logic of using discrimination to assess (pre-categorical) auditory encoding of continuous cues (Gerrits and Schouten, 2004; Massaro and Cohen, 1983; Pisoni and Tash, 1974). This work has cast doubt on the interpretation of the initial results favoring CP. Many of the original studies establishing CP (Liberman *et al.*, 1957) tested discrimination using an ABX procedure, in which two different stimuli are played in sequence (A + B), followed by a third (X) that matches the first or second. The participant's task is to say which of the first two matched the third. In this task, inability to discriminate the sounds should result in chance-level performance. This was found when stimuli lay within the same phonetic category, even as discrimination was strong for tokens that span the boundary (though few studies found true chance-level performance within categories).

Since the ABX task could be done without any labeling on the basis of the auditory percept alone, these results were interpreted as evidence that lower-level perceptual encoding is based largely on categorical information. However, deeper consideration of the task suggested that the ABX task may not isolate effects of perceptual discrimination from labeling (Gerrits and Schouten, 2004; Pisoni and Tash, 1974; Schouten *et al.*, 2003). The pattern of chance-level performance within a category depends on the exact nature of the discrimination task, and there is substantial evidence for within-category discrimination in several tasks. Pisoni and Tash (1974) measured reaction time in an AX same-different judgment task and found that within-category judgments of non-acoustically-identical stimuli elicited slower RTs than did identical stimuli, suggesting that the within-category differences were indeed perceived. Critically, they demonstrated that a distinct second stage of processing between encoding and categorization explains discrepancies between measures and suggested that the original ABX tasks were sensitive to both auditory encoding and later categorization processes. Similarly, Carney *et al.* (1977) showed substantial within-category discrimination using an oddball task.

Gerrits and Schouten (2004) more systematically explored the basis of the poor within-category discrimination in the ABX task. They reasoned that the ABX task relies heavily on working memory representations for the discrimination judgment. The listener is expected to compare the X stimulus with the A and B stimuli; thus, representations for A and B must be held in memory. However, memory traces fade with time, and auditory representations may fade faster and require more resources than lower dimensional categorical codes (e.g., the phonetic label). Consequently, in some conditions, listeners are forced to make discrimination judgments largely on the basis of the category labels they have assigned to those stimuli. Schouten *et al.* (2003) revealed that this memory demand is particularly problematic in the ABX task, as the reliance on memory leads to a strong bias to respond that the B stimulus matches the X stimulus.

To better explore the nature of discrimination, Schouten *et al.* (2003) systematically compared discrimination for the same stimuli in several discrimination tasks. Critically, the degree to which results matched the predictions of CP was a function of the memory demands for a task—tasks with lower memory demands revealed better within-category discrimination—and with the lowest memory-load task (what they term a 2IFC task), discrimination was strong and completely unrelated to the category boundary. A particularly notable finding was that with increasing inter-stimulus intervals (e.g., between the successive stimuli in the discrimination task), performance became more categorical—again implying the decay of memory representations is a critical feature.

Even more compellingly, Roberson *et al.* (2009) showed that CP effects for non-linguistic stimuli (color perception) could be completely eliminated when memory demands were removed from the task. In this case, because the stimuli do not have any inherent temporal component (unlike speech stimuli), they could be presented simultaneously. This presentation revealed a completely flat discrimination function, suggesting that CP effects for color categories are driven by the working memory demands of prior discrimination tasks.

These data present a damning view of the initial interpretation of discrimination data in support of CP: in these studies, *poor within-category discrimination likely reflected what was encoded in working memory and/or how stimuli were categorized rather than auditory encoding itself*. The closer a task gets to directly assessing the encoding of the perceptual cues, the more evidence there is that auditory encoding is not fully predicted by category identity. The prior evidence of CP in discrimination tasks seems to be an artifact of the way these tasks measure sensory encoding of speech.

IV. TOWARD A MORE GRADIENT MODEL OF PERCEPTION

The conclusion from this work appears to be that pre-categorical auditory encoding maintains continuous acoustic detail. This challenges one of the hallmarks of the strongest form of CP: that perceptual discrimination reflects categories themselves, without access to subcategorical detail. The broader theoretical assumption of sharp categorization has been more directly challenged by an array of tasks using different methods [e.g., priming (Andruski *et al.*, 1994), rating scales (Allen and Miller, 2004; Massaro and Cohen, 1983; Miller and Volaitis, 1989); eye-tracking (Kapnoula *et al.*, 2021; Kapnoula and McMurray, 2021; McMurray *et al.*, 2002, 2009; Ou *et al.*, 2021); and electroencephalogram (EEG) (Kapnoula and McMurray, 2021; Sarrett *et al.*, 2020; Toscano *et al.*, 2010)]. These tasks provide converging evidence that within-category acoustic differences can persist through response generation for some tasks. This has been interpreted as evidence that categorization is not discrete, but highly graded. For example, VOTs that are close to prototype values (e.g., 50 ms for a /p/) are categorized more robustly than less canonical VOTs (e.g., 30 ms) that are still in the same category. Later empirical and theoretical work has argued that this may in fact be beneficial for categorization and word recognition by helping listeners to be more flexible in making decisions in the face of uncertainty (Clayards *et al.*, 2008) and to more easily revise initially incorrect decisions (Kapnoula *et al.*, 2021; McMurray *et al.*, 2009). This challenges the notion that encoding discards continuous detail—even at the category level, sensitivity to fine-grained detail leads listeners to activate phoneme categories gradiently. Thus, continuous information persists throughout speech categorization.

V. ET 2, AFC?

The mounting evidence against theoretical claims of CP and the growing understanding of the nature of discrimination tasks have substantially curtailed their use as straightforward indices of speech perception. However, identification tasks of the sort used to support CP remain commonplace and have not received the same thorough analysis that discrimination tasks have undergone. This task is worth revisiting.

In a typical forced-choice identification task, participants hear stimuli that range between two or more prototypes and then classify each stimulus as a member of one or the other category in a forced choice between the discrete category responses [*n*-alternative forced choice (*n*AFC); note that here we focus primarily on 2AFC, but these concerns apply to all forced-choice tasks]. On its surface, this type of task seems an ideal way to assess speech categorization: it is simple and directly captures listeners' identification of the speech stimulus. Indeed, this task has clear, uncontroversial applications in studies that assess changes in category boundaries, like in perceptual learning (Kraljic and Samuel, 2005; Kraljic *et al.*, 2008; Norris *et al.*, 2000), accent adaptation (Reinisch and Holt, 2014; Sumner, 2011), talker normalization (Johnson *et al.*, 1999; Strand and Johnson, 1996), and context effects (Coady *et al.*, 2007; Holt, 2006). Moreover, the fairly straightforward nature of the task makes it amenable to use with diverse populations [e.g., young children (Hazan and Barrett, 2000; Slawinski and Fitzgerald, 1998), people with language impairments (Robertson *et al.*, 2009; Serniclaes, 2006; Sussman, 1993), and non-native speakers (Aoyama *et al.*, 2004; Goriot *et al.*, 2020; Sebastián-Gallés, 2011; Sebastián-Gallés and Bosch, 2002)]. In these applications, the relevant measure is the degree to which the estimated boundary moves due to learning, context, etc.

However, interpretation of other properties of the identification curve (e.g., the slope and separation of the asymptotes or amplitude) may not be as straightforward. CP was premised on the idea that listeners categorize a stimulus into discrete options at an early perceptual stage, and this in turn gives rise to the step-like response function. This made interpretation of the slope fairly transparent: a departure from a steep slope clearly derives from some kind of suboptimal (i.e., not step-like) categorization process. However, with the fall of CP, this underlying model cannot be assumed, and we now believe that the underlying categorization function is not step-like at all and is better characterized by a gradient transition between phonemes. As we argue next, this creates a large interpretive ambiguity for forced-choice identification tasks that show CP-like patterns. Critically, why does identification seem categorical if perception is not?

These challenges seem to be particularly relevant for studies of individual differences [e.g., second language (L2) learners] or clinical populations (e.g., dyslexia), where forced-choice identification is regularly used as an index of speech categorization ability. In these paradigms, categorical response patterns are thought to reflect "strong" ability (Coady *et al.*, 2007; Robertson *et al.*, 2009; Serniclaes, 2006; Sussman, 1993). Here, the focus is typically on the slope of the categorization function—how sharply do listeners shift from primarily one response to another. This interpretation (premised on CP) relies on this assumption of optimal responding. A steep slope is thought to reflect a well-defined, sharp perceptual category boundary—that is, effective CP at the level of initial auditory encoding [Fig. 2(A)]. Meanwhile, weak categorical perception would be indicated by a shallower slope, with the mean of responses in a middle region for tokens closer to the category boundary [Fig. 2(B)]. Note that in this context, a shallow slope reflects a "deficit in categorical precision" (Serniclaes, 2006), or "weak categorization" (Robertson *et al.*, 2009), given that CP is thought to be the typical and optimal response pattern. That is, according to CP, even though all participants are underlyingly categorical, some may have noisy auditory encoding or cue-to-category mapping. As a result, stimuli close to the boundary could sometimes be errantly perceived as the wrong category, but they are still thought to be perceived categorically (that is, they are "fully" or completely in the wrong category). This noisy encoding leads them to look gradient in the aggregate.

There are two related problems with this interpretation. First, it assumes that the CP-derived pattern of responding is optimal and therefore any deviation from sharp categorization must be suboptimal. Yet/however, if the activation of speech categories is gradient (and this is beneficial for listeners), is this the case? Shouldn't listeners be striving for a shallower (more gradient) slope? Second, this interpretation of the slope as an index of the quality or nature of speech categorization rests on the assumption that the proportion of responses is a veridical index of the underlying auditory encoding and subsequent categorization of the acoustic information. That is, the assumed model suggests that a choice of one of the two responses is a straightforward probabilistic readout of initial encoding, where a single category identity is perceived. In fact, as we argue here, if there is

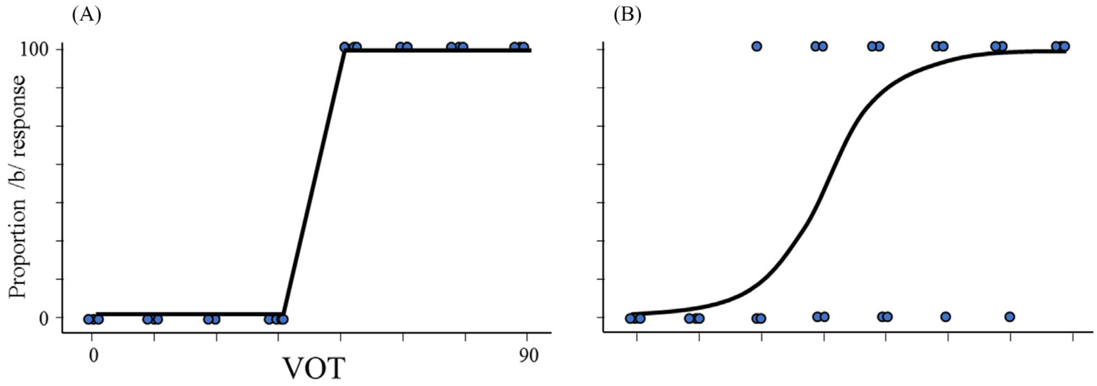


FIG. 2. (Color online) Response profiles in the 2AFC task. The black curve signifies the mean responses, and the blue dots signify individual responses on each trial. (A) A categorical pattern, with highly consistent responses to each item in the continuum. (B) A less categorical pattern, with increasing inconsistency to items close to the category boundary.

uncertainty about (1) the degree of noise in the encoding of cues like VOT, (2) the underlying category structure, and (3) the mapping between categories and responses (as there appears to be for special populations), then the 2AFC slope is ambiguous: both steep and shallow slopes can arise from both categorical and gradient encoding.

A. Steep slopes can mean different things

A steep slope is classically taken as one piece of evidence for CP; the listener is thought to discard irrelevant (within-category) stimulus variability. This makes sense if a listener's speech categorization maps cleanly onto the 2AFC response space (e.g., phonemes like /b/ or /p/) [Fig. 3(A)]. In this scenario, steep slopes emerge because a listener has an underlyingly categorical encoding of the perceptual cues, which is then mapped directly onto the two responses. That is, listeners "hear" the stimulus as /b/ or /p/ and then respond accordingly. This is the assumption in classic forms of CP. Specifically, the idea is that categorization is already accomplished during auditory encoding (or encoding is so substantially warped by categorization that category identity is the primary determiner of encoding), such that the mapping from perception to categorization is essentially a copy process. In this assumption, whether we measure at encoding or at categorization is irrelevant, as both stages are predicted to be categorical.

Alternatively, the two stages could operate more independently but still produce a pattern of responses with a steep slope between categories. In this case, listeners might hear the fine-grained differences within a category (e.g., Toscano and McMurray, 2010) but still impose a fairly rigid threshold when mapping cue values to categories: a VOT below 20 ms *always* maps to a /b/ [Fig. 3(B)]. In this case, the initial encoding is continuous, but the mapping from cues to categories discards the continuous information.

Yet another alternative could yield steep categorization slopes despite gradient category mapping. In this case, the mapping from cues to categories could also itself be graded (e.g., Miller, 1997) [Fig. 3(C)], but the nature of the 2AFC task could force the participant to respond with a dichotomous choice despite underlyingly graded activation of categories. This assumes some kind of intermediate response-mapping process to convert a graded category judgment to a discrete response in the particular experiment (a linking hypothesis).[2] One could obtain a categorical function if this category-to-response-mapping process takes the form of a winner-take-all mechanism in which the listener always chooses the response that is most probable [which Nearey and Hogan (1986) argue is the optimal response function]. For example, if a stimulus that is 60% /b/-like leads to a /b/ response every time, then any gradations in the underlying category mapping would be lost during response generation, and we would expect the same steep slope as assumed for CP. Critically, this latter case produces steep categorization because of the process of response generation rather than from perceptual or categorization processes—at every level, the speech processing is gradient! Given the strong evidence from Miller's and Massaro's work (Massaro and Cohen, 1983; Miller, 1997) for a gradient prototype-like structure to categories, some version of this must be true to account for sharp categorization results in forced-choice tasks. This process of winner-take-all categorization would result in the same steep categorization function described for the classic CP pattern, but it arises at a different point in the process (Nearey and Hogan, 1986).

Thus, a steep slope can emerge from at least three different situations, two of which include some element of gradiency in the system.

B. Shallow slopes can mean different things

The same interpretive difficulties arise for shallow slopes. The classic interpretation is that a shallow slope reflects noise at the level of auditory encoding. A listener might be truly categorical but have noise in their encoding of cues like VOT [Fig. 3(D)]. That is, on some trials, a VOT of 15 (a /b/) is actually heard as 25 ms (a /p/). In this scenario, even if the listener has a discrete boundary, on some trials, the noise in the cue encoding would cause the categorization response to "flip" due to errant encoding. Note that this same noise could not make a large impact near the end points—if a 50 ms VOT is heard as 40 ms, it would still be a

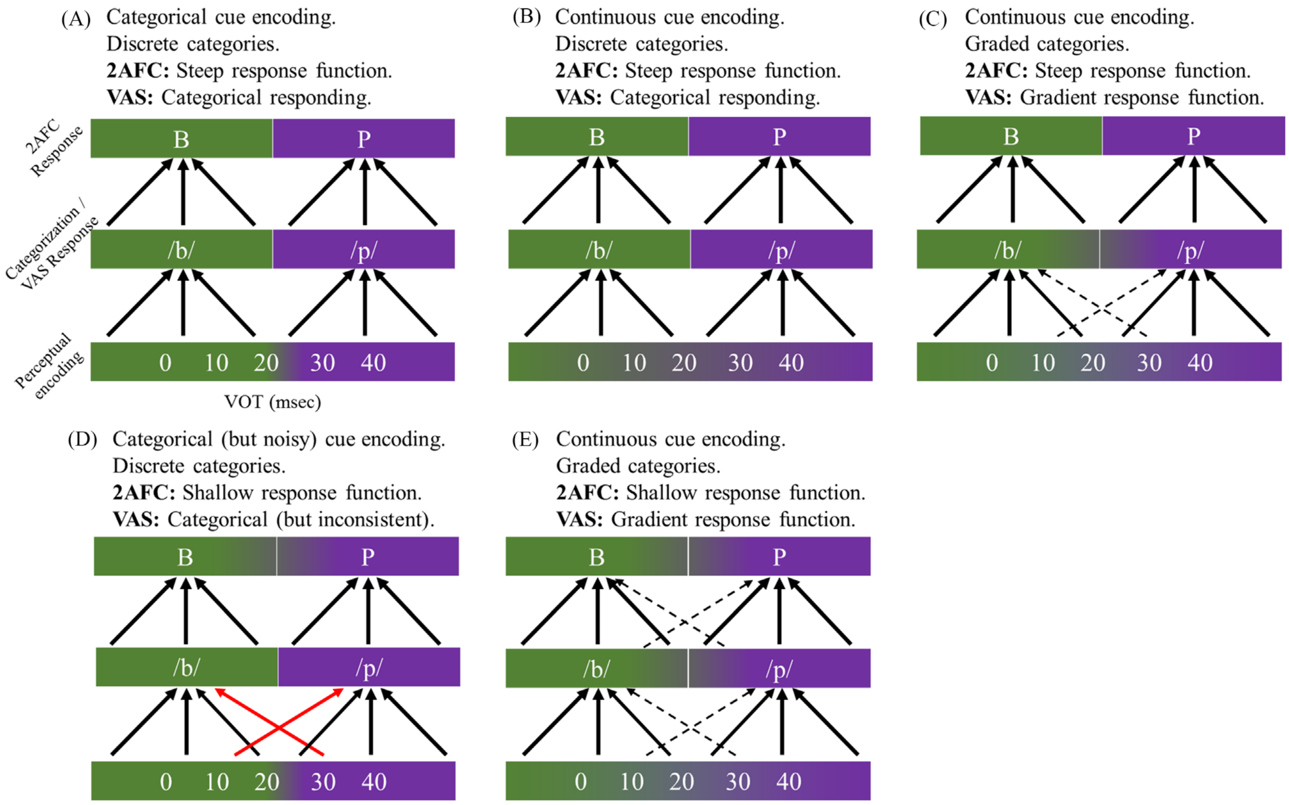

(A) Categorical cue encoding.
Discrete categories.
2AFC: Steep response function.
VAS: Categorical responding.
(B) Continuous cue encoding.
Discrete categories.
2AFC: Steep response function.
VAS: Categorical responding.
(C) Continuous cue encoding.
Graded categories.
2AFC: Steep response function.
VAS: Gradient response function.
2AFC Response
Categorization / VAS Response
Perceptual encoding
B
P
/b/
/p/
0 10 20 30 40
VOT (msec)
(D) Categorical (but noisy) cue encoding.
Discrete categories.
2AFC: Shallow response function.
VAS: Categorical (but inconsistent).
(E) Continuous cue encoding.
Graded categories.
2AFC: Shallow response function.
VAS: Gradient response function.

FIG. 3. (Color online) Potential mappings between encoding, categorization, and responding in the 2AFC task and the VAS task. This includes an initial encoding process, in which more continuous encoding is signaled by a color gradient, and more categorical encoding is signaled by more dichotomous color transitions; a categorization process of mapping to some form of speech categories (here identified as phonemes), which is also the expected mapping to the continuous VAS responses; and a 2AFC response process when binary response options are given, with similar color coding. (A) The classic CP assumptions, with clear categorical encoding that leads to discrete categories and unambiguous discrete responding. (B) A pattern in which continuous cue encoding maps to discrete categories via a winner-take-all mapping. This function leads to discrete responses that look like CP. (C) A pattern in which gradient cue encoding maps onto gradient categories, but then a winner-take-all response function leads to discrete responses that mirror CP patterns. (D) A pattern in which categorical encoding includes encoding noise, such that stimuli are sometimes misencoded. This maps onto discrete categories, but the noise at encoding means that stimuli near the category boundary are more likely to elicit the wrong category response. This leads to a shallow slope. (E) A pattern in which continuous encoding maps onto graded categories, which are then mapped to responses using probability matching. This leads to a shallow slope.

/p/. As such, noise in perceptual encoding would lead to more errant responses near the boundary. Aggregated across repetitions, this would produce a shallow slope. Supporting this, Ou and Yu (2022) constructed neural representational dissimilarity matrices and found that listeners with steeper 2AFC slopes showed more veridical speech encoding in subcortical areas (higher similarity between acoustic and subcortical representations), supporting the idea that the 2AFC slope reflects perceptual noise.

However, a shallow slope could also emerge because of gradient encoding and categorization. In this case [Fig. 3(E)], listeners maintain continuous detail throughout the system and then do their best to map this continuous detail onto the response space. This could occur if listeners use a form of probability matching to map their underlying categorization onto a response (Clayards *et al.*, 2008)—if a listener perceives a token as 60% /b/ but 40% /p/, they might identify this token as /b/ on 60% of the trials. This again would result in a shallow slope for the categorization function, which is consistent with good perceptual encoding and gradient categorization. This interpretation of the 2AFC is consistent with the graded nature of speech perception.

Thus, a shallow slope in a 2AFC task could be consistent with either categorical cue-to-category mapping or a gradient one.

C. The 2AFC's slope tells us nothing about the underlying perceptual encoding

As we have described, categorical patterns can arise despite gradient encoding and even gradient categorization, and gradient patterns (shallow slopes) can emerge despite CP. Yet other patterns are also possible—for example, noise in the response function that maps categories to responses could affect the shape of the categorization function completely independent of encoding *and* categorization. Depending on the assumed model of categorization, this response-level noise could affect the slope of the function or the degree of separation of the asymptotes (a common observation in studies of impaired listeners). Indeed, some populations are likely to have noisier response mappings—for example, younger children are notoriously noisy in their responses (Vane and Motta, 1980). This population-level difference is hugely problematic for attempts to compare speech perception ability between groups [e.g., developmental studies (Hazan and Barrett, 2000;

McQueen *et al.*, 2012; Slawinski and Fitzgerald, 1998) and comparing typical populations to those with language impairment (Blomert and Mitterer, 2004; Noordenbos *et al.*, 2012)]; what seems to be evidence of group-level differences in how speech is perceived and categorized might instead reflect differences in how the groups map categories onto the forced-choice response space.

The uncertainty as to how an underlying percept is mapped to categories and then to a forced choice makes it difficult to divorce differences in the encoding of cues from differences in how they map to categories and then to responses. The slope of a 2AFC categorization function bears no necessary relationship to a listener's underlying auditory encoding—noise in encoding or response generation could make them appear more gradient than they really are, or the forced choice of either /b/ or /p/ could introduce a need to convert continuous category representations into discrete response options. Thus, the 2AFC task lacks discriminant validity and measurement specificity as a tool for measuring whether speech categorization is graded or categorical. Differences between populations (or even between listeners) could arise at several levels, and the same pattern across populations could arise from different constellations of underlying processes.

A key source of these concerns is the fact that the slope of the function is based entirely on binary responses that are averaged across repetitions. A shallow slope can only emerge if a participant provides variable responses to the same token across repetitions (e.g., selecting /b/ on 60% of repetitions of a given stimulus and /p/ on 40%). Thus, participants cannot report a gradient percept for an individual trial—on each trial, they are required to identify the token as /b/ or /p/ (for example), even if their category representation fell somewhere between the two. A listener can only demonstrate gradiency in the aggregate across multiple repetitions for a stimulus. However, because this pattern emerges only because of inconsistent responses to stimuli, it might signal encoding noise instead of true gradiency. This is why the nature of the response mapping (winner-take-all vs probability matching) can play such a big role in deriving predictions for a 2AFC task. In contrast, if there was a way to more directly estimate how listeners categorize stimuli on individual trials, one could in principle separate these factors by looking at the consistency of the individual responses as well as the aggregate patterns of responses across stimuli and repetitions.

This ambiguity in the factors that elicit slope differences in 2AFC tasks could lead to dramatic misinterpretations of speech categorization, particularly in studies comparing between groups. For example, in the case of development, younger children often show shallower slopes (Hazan and Barrett, 2000; Slawinski and Fitzgerald, 1998), but this cannot tell us if shallower slopes are due to higher noise [Fig. 3(D)] or higher gradiency [Fig. 3(E)]. McMurray *et al.* (2018) directly divorced these two using eye-tracking. They found that younger children had shallower slopes than older children when measured by forced-choice identification, consistent with most prior work. However, these same younger children showed substantially *less* sensitivity to within-category detail in the pattern of eye movements to lexical competitors. This is analogous to being *more* categorical! That is, their 2AFC slopes do not reflect over-sensitivity to fine-grained detail, but rather noise in the system (and increased sensitivity is obtained when this is overcome later in development).

Similar conflicts between encoding, categorization, and response mapping could arise for other population comparisons. For example, L2 learners and people with language disorders might have gradient response profiles, while steep response functions are assumed to be the end goal—i.e., these listeners are supposed to be working toward CP. Much bilingualism research has relied on this assumption and consequently relied on problematic discrimination tasks to assess L2 language ability (Aoyama *et al.*, 2004; Goriot *et al.*, 2020; Sebastián-Gallés, 2011; Sebastián-Gallés and Bosch, 2002; Werker and Tees, 1987). As a result, other factors such as the learners' age of acquisition and proficiency are often interpreted to reflect "non-native" discrimination skills (Bosch, 2010). However, the increasing evidence that skilled listeners maintain continuous detail through categorization casts doubt that CP should be the end goal for these populations. In fact, for a bilingual listener confronted with multiple overlapping phonological systems, it may be more optimal to maintain a gradient mapping between cues and categories to permit more flexibility. Instead, it is possible that increasing steepness in their responses on 2AFC tasks might reflect changes in how they map categories to responses [e.g., the distinction between Figs. 3(C) and 3(E)] rather than changes in their encoding and that the more gradient functions of early L2 learners could be an adaptive response to uncertainty. Without a clear mechanistic definition of what constitutes successful categorization, it is unclear how to interpret these results.

VI. ALTERNATIVE PARADIGMS

Several alternative approaches have been used to provide a more nuanced view of the nature of speech categorization, and these have often produced a different view than the standard identification tasks. These tasks overcome some of the difficulties with the 2AFC and may allow us to isolate different aspects of the system. However, they all face theoretical challenges to their interpretation and/or practical challenges to their use.

One simple way to bypass the purely dichotomous nature of 2AFC is to ask listeners to *rate the goodness* of individual tokens (Allen and Miller, 1999; Massaro and Cohen, 1983; Miller and Volaitis, 1989). A listener might label two tokens as both /b/ with high consistency but still recognize that one is more prototypical of the /b/ category than another. Such a case would predict a categorical pattern of identification responses despite more continuous underlying categorization [e.g., Fig. 3(E)]. Miller and colleagues have used such tasks across several studies and find that listeners' ratings show a great deal of sensitivity to within-category acoustic detail, as listeners judge tokens closer to the category boundary to be less good exemplars of the category (Miller and Volaitis, 1989). This suggests that listeners are able to retain the within-category acoustic detail when the task is structured appropriately to assess it.

Other studies use *priming* to indirectly assess whether different tokens within a category activate words to varying degrees (Andruski *et al.*, 1994). Under CP, within-category acoustic variation should not impact word recognition—activation of the word *king* should be identical whether the VOT of the onset /k/ is highly prototypical or closer to the boundary with /g/, as long as it is on the /k/ side. However, the opposite is attested; Andruski and colleagues found less priming for a semantically related word after hearing a word closer to a category boundary—*king* with a less prototypical /k/ produces less priming of *queen* than it does with a prototypical /k/ [see also Utman *et al.* (2000)]. This suggests that listeners are sensitive to these minute within-category differences and that this subcategorical information cascades through to lexical-level processing.

One concern with these tasks is that they use fairly indirect means and metalinguistic judgments to assess speech categorization. The goodness-rating task asks a listener to make a subjective judgment of the quality of a stimulus, which might entail processes outside of speech categorization. Their subjective interpretation of the instructions might lead them to judge a stimulus as "bad" based on acoustic but non-phonetic criteria, such as voice quality. That is, a listener might exhibit CP within the speech system but still retain continuous acoustic detail for non-phonemic judgments of things like voice quality. In addition, this task dissociates with speech categorization for some stimuli—a token that is non-prototypical but far from the category boundary (e.g., a heavily pre-voiced /b/) might be rated as low on goodness but would be unlikely to garner any perceptual support as a /p/ (or any other non-/b/ phoneme). On the other hand, the priming task relies on indirect sequelae of within-category variation: the speed of lexical judgments of another word presented after the item of interest. This requires a chain of several assumptions to link the behavior to speech perception. In addition, there is no measurement of how the listener actually identifies the prime token. Variations in the mean lexical decision times after the different primes might arise because of variations in listeners' category boundaries—listeners might identify the prime as *ging* on some trials and, thus, produce less priming despite perceiving categorically.

More direct approaches attempt to assess how the listener processes speech in real time, using non-dichotomous measures. Eye-tracking in the VWP has shown that the dynamics of fixations to items vary as a function of within-category acoustic differences (Dahan *et al.*, 2001; McMurray *et al.*, 2003, 2009; Salverda *et al.*, 2003). In a variant of this task used to study speech categorization, a participant hears tokens from a speech continuum (e.g., *beach-peach*) and sees a display that includes pictures of each end point, as well as unrelated foils (McMurray *et al.*, 2003). Eye movements to the pictures are measured as they hear each token, and the proportion of trials in which a listener fixates a given target across time is thought to reflect the degree to which a word is activated (Allopenna *et al.*, 1998). In this version, the VWP allows a dissociation between the final response and the way that words were activated before the listener generated this response because analysis can be conditioned on the item that was ultimately selected (e.g., only consider trials where the listener clicked *beach* as the target) to see whether the pattern of looks preceding a *beach* response varies for different acoustic tokens. This can directly disentangle the response from the processes preceding it by (at least in the aggregate) avoiding averaging trials with different categorical judgments. Consider a situation in which on some trials a listener discretely heard a VOT of 10 ms (normally a /b/) as a /p/. They then fixated it, bumping up the average. However, on these trials, they should also click on the /p/, and the trial would be discarded. Thus, the inferential logic of conditioning the analysis of eye movements on the response avoids some kinds of averaging artifacts.

This approach has provided strong evidence that listeners encode and use subcategorical acoustic detail despite categorical identification (McMurray *et al.*, 2003; McMurray *et al.*, 2009). As mentioned above, McMurray *et al.* (2018) used this approach and found shallow identification slopes for young children but more categorical patterns of eye movements. This pattern highlights the dissociation between perception and response and reinforces the need to consider the multiple processes that contribute to responses in speech tasks in tandem.

However, the VWP also faces interpretive hurdles. On the practical level, this approach requires a large number of trials to gather interpretable data. This can limit the applicability of this approach for some questions—long studies with lots of repetition are not always feasible. In addition, high numbers of repetitions could also affect processing strategies; as listeners become more attuned to the specific stimuli, they may begin to focus on aspects they would not in a typical, non-experimental speech context. More importantly, at any given point in time, a participant is either looking at a target picture (e.g., of the *beach*), or they are not—the measure is binary. The interpretation that listeners process speech gradiently arises from aggregation of many time points over several trials; the proportion of trials when the *beach* is fixated is lower for /b/s close to the category boundary than prototypical /b/s, but on any individual trial, at any given time, they either are—or are not—looking at *beach*. Although many of the averaging concerns are handled by controlling for the ultimate identification (they clicked *beach* in each case), it is nonetheless tricky to argue for a gradient mapping from a binary response. For example, though it is implausible, some have argued that listeners might always have a discrete interpretation at any given point in time but vacillate between them over time *within a trial*. When fixations over different time points are averaged, it could make it look like a graded percept [see, for example, Spivey *et al.* (2005), who introduced this concern and ruled it out using mouse-tracking].

A related approach uses ERPs (e.g., Toscano *et al.*, 2010). ERPs are measurements of brain activity that are time-locked to a presented stimulus, so they can be used to index whether and when a listener shows differences in processing sub-categorically different stimuli. This method has been used to show that initial acoustic processing in the N1 component, approximately 100 ms after stimulus onset, reflects continuous acoustic information, independent of category identity (Getz and Toscano, 2021; Toscano *et al.*, 2010). More recently, Sarrett and colleagues leveraged this approach to show clear electrophysiological evidence for maintenance of continuous acoustic information during later processing stages up to 900 ms post-stimulus (Sarrett *et al.*, 2020). This confirms the argument from discrimination tasks that the auditory encoding of acoustic/phonetic cues is continuous and that graded representations persist for substantial periods during processing.

Similar ERP components have also been used to examine the gradiency of categorization following a similar logic to the VWP. In this case, the P3 (which is elicited ∼300–800 ms post-stimulus) is an index of perceived category prototypicality. That is, if listeners are tracking fine-grained detail, the P3 should be strongest at 0 ms and get smaller as the VOT approaches the boundary. Critically, like in the VWP, analysis can be predicated on the ultimate response, thus ruling out some types of averaging effects (Kapnoula and McMurray, 2021; Toscano *et al.*, 2010).

However, ERPs still face many of the same implementation and interpretation issues as the VWP. They depend on a very large number of trials to overcome trial-level noise. Moreover, even though the measure is fundamentally continuous, individual trials are uninterpretable, making it difficult to use the trial-by-trial variation to make inferences about the system. Finally, the use of averaging in ERP leaves us in the same conundrum as other averaging techniques. Although we can condition analysis on the explicit response, it is possible that things that look gradient might reflect different subsets of trials in which participants briefly entertained the alternative categorical interpretation.

Moreover, paradigms like the VWP and the P3 in which analysis of each trial is predicated on a given response have another problem: they limit interpretation to only half of the continuum at a time. This is also true in phoneme goodness ratings; the dimensions that constitute a good /p/ in general may not be the same as the ones that contrast a /p/ from a /b/ or a /p/ from a /t/. The implicit contrast afforded by a 2AFC task serves to highlight the dimensions that are most relevant in the task.

The VWP and ERP studies, in fact, do highlight this contrast but still do not allow us to analyze the full continuum in a satisfactory way. In the VWP, we analyze looks to the *peach* for all trials in which the participant signaled that they heard *beach*; the P3 is analyzed only for trials in which they said the target was present; and in the prototypicality work, listeners are explicitly told to rate how good of a /p/ the stimulus was. This only gives us information about perception of one side of the contrast, making the task something of a 1AFC. This, in turn, makes it difficult to understand what is going on toward the center of the continuum, where participants might be less confident in the category identity of the stimuli. In fact, the confidence in the eye-movement record or ERP record goes down in exactly these portions of the continuum as these are the regions in which the listener is more likely to choose the other category (and, therefore, the regions in which more trials are discarded).

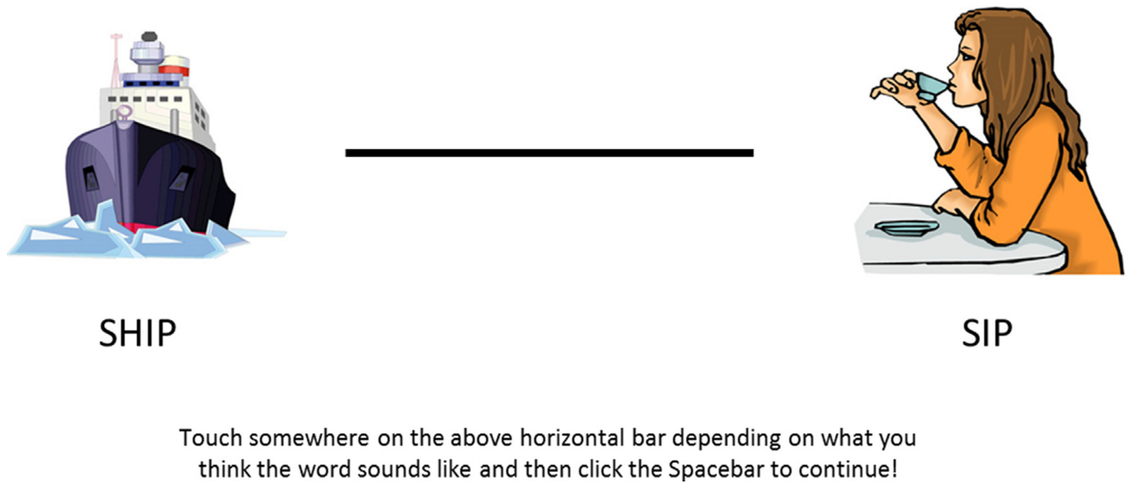
This limitation of VWP/ERP tasks in measuring categorization close to the boundary appears particularly problematic in light of recent evidence showing that differences in categorization gradiency are largely due to differences in the degree of perceptual warping of acoustic cues close to the boundary (Kapnoula and McMurray, 2021). That is, it is difficult to measure precisely the region that may be of most interest. The classic 2AFC task had the advantage of direct comparison of all trials no matter what response was chosen. We thus need techniques that also have this advantage but that also avoid the hazards of forced-choice tasks.

VII. A PROMISING ALTERNATIVE: THE VISUAL ANALOG SCALE (VAS)

A promising alternative task embraces the continuous decisions of Miller's rating scales but focuses responses on phonetic judgments along a continuum. We call this task the VAS task[3] [Kapnoula *et al.*, 2017; Kong and Edwards, 2016; Munson and Carlson, 2016; see also Massaro and Cohen (1983)]. The VAS capitalizes on the benefits of tasks like 2AFC, which directly ask listeners about their phonetic/lexical interpretations of stimuli in a context that explicitly highlights the relevant dimension (by contrasting two categories). At the same time, the continuous task allows participants to report more continuous representations for a single trial.

In the VAS task, listeners hear tokens along a speech continuum (e.g., /b/–/p/) and respond by marking on a line how closely they correspond to either end point.[4] This line represents a scale between these end points—the listener can signify if the token is a perfect /b/, a perfect /p/, or somewhere in between. Figure 4 displays a version of this task used in developmental research by several of the authors of this paper as part of the Growing Words Project.

The VAS task removes the need for listeners to convert a potentially continuous categorization space into a dichotomous response space. In Fig. 3, the VAS would be expected to allow a more direct mapping between continuous levels of category activation and the response space, rather than requiring dichotomous responding. Additionally, it removes the uncertainty of forced-choice tasks as to whether listeners probability-match or use a winner-take-all decision rule. This is because gradiency can be assessed within individual trials, instead of via the average response across trials. In this way, the task can better differentiate between gradient categorization and variable responding. Last, in contrast to



Touch somewhere on the above horizontal bar depending on what you think the word sounds like and then click the Spacebar to continue!

FIG. 4. (Color online) A schematic of the VAS response display as used for a developmental population. The bar in the middle is compressed here for visualization purposes; in practice, it would extend across the length allowed by the monitor. When a participant makes a response by touching or clicking along the bar, their response is marked by a line that appears at that location.

single-category goodness ratings, VAS better reflects speech categorization as the two anchor categories focus the listeners' attention on the phonetically relevant dimension(s).

A. Indices of categorization[5]

In some ways, the VAS functions similarly to a 2AFC task. Ratings of stimuli can be averaged and plotted akin to the 2AFC task to assess things like *slope* and *category boundary* [Fig. 5(A)]. Unlike the 2AFC task, however, listeners need not be strongly biased toward 0 and 1 and even at the end points may show an asymptote that does not reach either end point. This raises two additional indices that are not commonly considered in 2AFC work: the *amplitude* (the distance between the asymptotes of the function), and the *bias* (if there is an overall bias to respond to one end of the continuum, regardless of step). All four of these indices are similar to those that might be used in the 2AFC task. However, there are two key advantages to VAS.

Note that amplitude (and closely related bias) has not always been included in psychometric approaches for 2AFC, in part because of an (often untested) assumption that forced-choice responding will lead to unambiguous responses to the end points. It is possible to fit 2AFC data (and also VAS data) with a two-parameter function that only estimates slope and crossover category boundary and assumes saturation at 0% and 100%. However, such an assumption can substantially alter slope estimates (Wichmann and Hill, 2001); a function with low amplitude but steep slopes may be estimated to have shallower slopes if amplitude is ignored. This is true for both 2AFC and VAS tasks.

First, in the VAS, the scores are not all 0 and 1, so effective estimates of categorization functions require substantially fewer trials. In principle, a single repetition of each token in the continuum could produce a continuous measure of perceptual judgment. In practice, reliable estimates of categorization require some stimulus repetition. Kong and Edwards (2016) achieved moderate test-retest reliability ($r = 0.48$) of gradiency estimates with a design including three repetitions of each stimulus. Three repetitions have proven sufficient to estimate categorization in several other studies as well (Kapnoula *et al.*, 2017; Kapnoula *et al.*, 2021; Kapnoula and McMurray, 2021). This is substantially fewer repetitions than are needed for a 2AFC task, in which gradiency can only be assessed by inconsistently responding to the same stimulus across repetitions. However, further research is needed to determine the optimal number of repetitions needed for reliability for different analytic approaches.

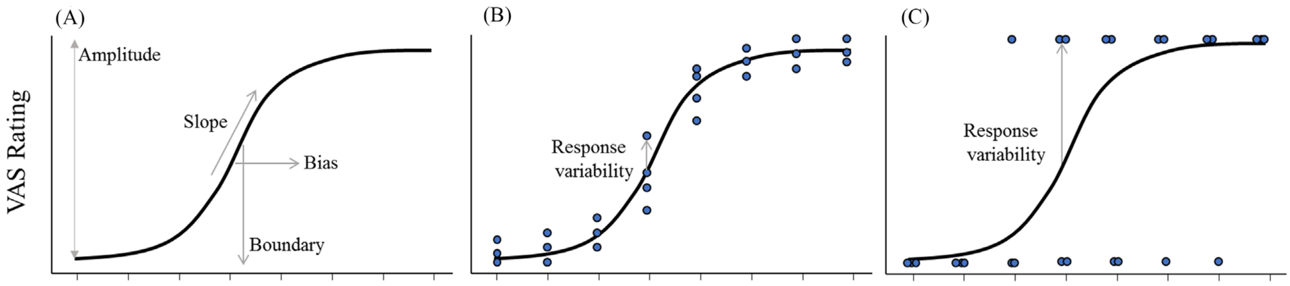


FIG. 5. (Color online) Indices of categorization in the VAS and possible patterns of data. (A) A schematic of averaged responses across a continuum, with identifiers of measures of amplitude, slope, and bias. (B) Individual responses overlaid on an average curve that shows a shallow slope that arises from low response variability. (C) An identical average curve that arises from high response variability because of more categorical, but inconsistent, responses.

Second, and more importantly, these measures are vastly more interpretable in the VAS task, where we can also estimate a fifth index: *response variability* [Figs. 5(B) and 5(C)]. Early work with the VAS (e.g., Kong and Edwards, 2016) visualized response variability with a histogram of the continuous responses (across all tokens). This allows us to ask if listeners tend to respond only with the end points or if they use the whole scale. However, a histogram defies an easy numerical index (which could be correlated with other measures). Moreover, their approach pools across continuum steps and, therefore, ignores whether or not listeners respond in a way that reflects the stimulus (e.g., the specific continuum step). That is, it reflects variability of responses across stimuli, not the variability with which a single stimulus is classified.

Later work (Kapnoula *et al.*, 2017) introduced a new way to examine response variability. Kapnoula computed the average response function for each participant (in this case using a nonlinear function, but it could also be just the mean at each step) and then computed the sum of the squared difference between each individual trial rating and the mean. This then reflects the averaged consistency or variability with which a listener responds to a specific stimulus (given the average response to that stimulus). If listeners' individual responses generally track the mean function, the sum of squared differences should be low, but if listeners have high variability (low consistency), the sum of squared differences should be high.

Response variability by itself may be a highly informative index of performance—particularly for children or clinical populations. However, its real value is when used alongside other metrics. By interpreting each of the traditional categorization indices in light of the response variability, we can achieve deeper insight into categorization and disentangle the kinds of ambiguities highlighted for the 2AFC task—particularly around slope and amplitude.

B. Slope

The VAS gives an estimate of the *slope* of the categorization function. This can be done over single cues, but it can also be calculated over multiple acoustic cues, allowing us to estimate gradiency at the phoneme level rather than the cue level. This is needed for categories that are defined by more than one cue (e.g., VOT and F_0 for the voicing contrasts). Supporting this, Kapnoula *et al.* (2017) present a novel nonlinear approach that can simultaneously compute the weighting of each cue and the slope over the diagonal boundary in a two-dimensional space [see McMurray (2017) for software].

The slope is often used as the primary measure of the degree to which speech categorization is discrete or gradient—participants with steeper slopes are thought to have more precise speech categories or to be more categorical than those with shallower slopes. However, when interpreted along with response variability, the VAS offers substantially more insight into the cause of slope differences.

For 2AFC, a shallow slope necessarily means that responses to specific stimuli were variable across repetitions—a token with a 60% /b/ score means that it was labeled as /b/ on 60% of trials and /p/ on 40% of trials. However, it is not clear whether this variability derives from noisy auditory encoding or from a gradient mapping (with a probability readout decision rule). In the VAS, this same pattern could arise for either reason as well, but they are now dissociable. First, a shallower slope on average could arise if the listener responds to the stimulus as 100% /b/-like on 60% of repetitions and 100% /p/-like on the other 40%—that is, if they responded identically to the way they would in 2AFC. However, this pattern could also arise if the token was marked at 60% of the way to the /b/ end of the scale on every trial. These two ways of achieving the same mean score likely reflect very different underlying processes. In the former, the participant appears to form a categorical percept each time they hear the stimulus but is inconsistent in which category is heard or which response is generated—the locus of the effect is noise in perceptual encoding. In the latter, the participant appears to form a graded percept and is highly consistent in responding—the locus is in gradient categorization.

Thus, a listener with a shallow slope, but who exhibits very consistent responding to stimuli along the scale, is one who exhibits precise encoding of subcategorical detail. In contrast, a listener with a shallow slope that arises from inconsistent responses that vary between the two end points of the scale might have more categorical encoding but less consistency in the encoding of acoustic cues. That is, they fit the model assumed by CP in which they are striving for a discrete mapping but suffer from noisy cue encoding. This distinction is important; a shallow slope alone could arise for very different reasons—which align with very different theoretical models of perception—and it is not enough on its own to understand the nature of that listener's speech categorization. We would not want to characterize these two participants as having identical speech categorization profiles; the VAS allows us to discriminate the two profiles, whereas they are confounded in 2AFC. Critically, it is the fact that response variability is computed relative to the mean (60%), not just in general, that allows us to separate these.

C. Amplitude

The *amplitude* of the function signifies the overall difference between the asymptotes of the response function (at the extremes of the continuum). A skilled listener responding to well-constructed stimuli should respond very close to one end of the scale for one end point of the continuum and very close to the other end for the other end point. However, often end points are not fully at 0% or 100% of the scale for either response for some tasks or for some populations (e.g., people with dyslexia; Manis *et al.*, 1997). In a 2AFC task, this means that participants sometimes choose the opposite response even for the least ambiguous responses, and this is usually interpreted as pure response noise or even guessing (and it is often described in psychometric functions as the "lapse rate"). However, in more modern theoretical views of speech, this could also derive from a situation in which speech perception is extremely gradient, biased to one category, or one in which the end points are not perceived as 100% representative (maybe competition between the two categories never fully resolved, or perhaps the listener does not have clear categories, such as in an L2 case). Under this view, if listeners adopt a probability-matching rule, this can then lead to inconsistent responses.

The VAS can differentiate between these sources of amplitude differences. If this pattern truly represents a lapse of attention or guessing, one might expect to see high response variability—on the subset of lapse trials, listeners' responses are all over the scale. In contrast, if listeners cannot fully separate the end points, we might expect to see low variability even as the amplitudes are not 0 or 1.

D. Category boundary

Similar to the 2AFC task, we can estimate a listener's category *boundary*, where judgments transition from one end point to the other. This boundary is often not of critical theoretical interest for studies of the speech categorization gradiency and is often factored out in analyses. However, it is crucial in studies of context effects, perceptual learning, accent adaptation, and the like, and it is straightforward to estimate in VAS. Consequently, switching to a VAS task yields no loss of information and provides a much richer and clearer view of other aspects of speech categorization. For example, a particular perceptual learning manipulation may shift the boundary, but it may also add uncertainty to the system, making the slope shallower (Clayards *et al.*, 2008; Theodore *et al.*, 2020). If this were measured with the VAS task, one could in principle make stronger claims about the source of the shallower slope.

E. Bias

The fourth property is the *bias*, which reflects whether responses are more likely to be to one side of the scale than the other. In a 2AFC framework, this would arise because a listener uses one response more often, regardless of the stimulus. This could be because of an overall bias to press that button independent of perception or because of a bias in categorization. In the VAS framework, these are somewhat differentiable. An overall bias could arise because listeners shift all responses toward one end of the continuum but still use the entire scale—a perceptual bias toward /b/. Alternatively, it could arise because the listener occasionally makes a discrete response of one end point, irrespective of the stimulus—sometimes they respond with /b/ no matter where on the continuum a stimulus falls. This is common, for example, in studies of brain-damaged patients (Dial *et al.*, 2019; Kocsis *et al.*, 2022). These two patterns of responding could both affect the overall bias, but they are dissociable using the VAS.

F. Permutations

The previously described indices are not independent and can be combined and permuted in various ways. Clearly, some indices are best interpretable in tandem—slope and response variability are much more meaningful as a combination when slope is low. Others are highly constrained: if the slope is high, response variability is necessarily low, and if the amplitude is near 1.0, bias cannot be other than 0.50. Thus, many of these may be more properly interpreted in a multi-dimensional space. There are also more targeted measures that may be possible. For example, it may be theoretically important to separate response variability at the asymptotes from variability in the boundary region. Finer grained analyses of the mean response at the boundary region may also be helpful for investigating warping; for example, one could compare the responses near the boundary to an idealized gradient or categorical function. Finally, RTs might offer further insight into the nature of the processes that lead to a response in the VAS, though the task may need to be modified to make RTs more reliable (for example, by encouraging rapid responding and not allowing listeners to change their responses). In sum, the flexibility permitted by the VAS task may enable a richer and more multi-dimensional picture of speech categorization than can be obtained with forced-choice tasks.

VIII. VALIDITY AND RELIABILITY OF THE VAS

Further investigation and development will improve the ability to link the VAS to underlying speech categorization processes. However, notable work has already demonstrated its validity and utility as a measure.

For example, Kong and Edwards (2016) asked whether participants' VAS responding is consistent across sessions. They tested adults on a /da/–/ta/ continuum, asking them to rate each stimulus on a continuous scale. Some participants preferred the end points of the line (i.e., their responding was more categorical), while others were more gradient, using the entire range of available responses. Importantly, participants performed this task (among others) twice, in two experimental sessions separated by about 1 week. The results showed significant correlations between the two sessions, both in terms of participants' overall use of the scale ($r = 0.48$) and in terms of how much they used the primary cue (VOT; $r = 0.67$). Note that other estimates of test-retest reliability are needed for other analytic approaches.

More recently, Kapnoula and McMurray (2021) assessed the task's validity, providing confirmation that the VAS task reflects sensitivity to within-category differences. The same participants performed a number of tasks, including a VAS task, a VWP task designed to assess speech categorization gradiency at the lexical level (McMurray *et al.*, 2002), and an EEG task designed to assess speech categorization gradiency at the level of speech categories (Toscano *et al.*, 2010). It was, thus, expected that the three measures would be related. Indeed, for both the VWP and the ERP task, the key effect (the slope of the fixations or P3 as a function of VOT) was related to the slope of the VAS responses. Listeners with more gradient VAS responding showed a larger effect of within-category differences on competitor fixations in the VWP and a larger drop off (within-category) in the P3 as the VOT approached the boundary. This pattern suggests that all three measures are tapping something fundamental about speech categorization [see also Kapnoula *et al.* (2021)].

As mentioned above, a shallow identification slope can reflect response variability, sensitivity to subcategorical differences, or a mixture of the two. Even though this applies to both 2AFC and VAS tasks, only the latter provides an independent measure of response variability. Taking advantage of this possibility, Kapnoula *et al.* (2017) examined the degree to which response variability was related to each type of identification slope. Specifically, to extract a measure of response variability, they first computed the difference between the predicted and actual rating for each VAS trial and then computed the standard deviation of the residuals. A set of correlational analyses showed that more consistent VAS raters had steeper 2AFC slopes (marginally significant), whereas no relationship was found between response variability and VAS slope (and the trend was in the opposite direction). These results suggest that the 2AFC slope is more likely to reflect noise in encoding or differences in category-response mapping rather than perceptual gradiency, whereas VAS slope is more likely to reflect gradient perception of speech categories. This idea is also in line with recent results by Fuhrmeister and Myers (2021), who used structural functional magnetic resonance imaging (fMRI) and found that speech gradiency and response variability (extracted from a VAS-type task) have distinct structural underpinnings.

IX. INITIAL SUCCESSES OF THE VAS AND PROMISING DIRECTIONS

Given the validity and reliability of the VAS measure of speech categorization gradiency, a number of recent studies have begun using it to ask more fundamental questions about why some people are more gradient than others and whether there are consequences for gradient categorization for spoken language processing.

Kapnoula and McMurray (2021) addressed the first question, by examining the N1 ERP component. As mentioned above, this component is thought to reflect the continuous encoding of acoustic/cue information (e.g., of VOT), independent of phoneme category identity (Toscano *et al.*, 2010). Kapnoula and McMurray (2021) used this component to ask whether individual differences in VAS responses reflect differences in the early perception of acoustic cues. Indeed, results showed that the linear relationship between N1 and VOT was disrupted in the perceptual space around the category boundary, but only for listeners who exhibited a more categorical pattern of VAS responding. This finding not only further validates the VAS task as a measure of speech processing, but it also points to early cue encoding as a source of gradiency. This finding also fits nicely with the pattern of results across experiments showing that correlations between VAS slopes for different continua depend on the acoustic similarity between stimuli sets (Kapnoula *et al.*, 2017; Kapnoula *et al.*, 2021). Together, these results suggest that individual differences in gradiency are most likely due to patterns of encoding of specific cues, rather than a global approach to speech categorization.

However, continuous encoding of a specific acoustic cue may not be a sufficient condition for the appearance of gradiency at the level of responses—there are small but reliable influences of domain general cognitive abilities. Specifically, Kapnoula *et al.* (2017) found that working memory (measured by an *n*-back task) was a significant predictor of VAS slope, with higher working memory scores predicting more gradient VAS responses [see also Kim *et al.* (2020)]. This finding points to a mediating role of working memory; even when cue encoding is gradient, the maintenance of this gradiency depends on working memory. Looking at a different aspect of executive function, Kapnoula and McMurray (2021) found that participants with better inhibitory control (measured by a spatial Stroop task) showed more gradient VAS responding. This finding again points to a modulatory role of inhibitory control; higher gradiency may lead to the parallel activation of multiple representations, but then better inhibitory control is needed to flexibly manage these activations and allow for gradiency to be reflected in the response. Together, these results suggest that speech gradiency stems from continuous encoding of acoustic cues but needs to be maintained to appear at the level of VAS responding.

Yet/however, what exactly is the functional role of gradiency in spoken language processing? For example, it has been suggested that gradient speech perception may allow for more flexible processing of the speech input (Clayards *et al.*, 2008; Kleinschmidt and Jaeger, 2015; McMurray *et al.*, 2002; Miller, 1997). The VAS can help us address this question too by using an individual differences approach. First, higher sensitivity to subcategorical detail (i.e., higher gradiency) may allow listeners to more flexibly combine different speech cues. Indeed, there are now several studies showing that higher gradiency is linked to higher use of secondary cues (at least for some contrasts; Kapnoula *et al.*, 2017; Kapnoula *et al.*, 2021; Kapnoula and McMurray, 2021; Kim *et al.*, 2020; Kong and Edwards, 2016).

Taking a different approach, Kapnoula *et al.* (2021) used a VAS task along with a VWP task aimed at measuring listeners' ability to recover from lexical garden paths (McMurray *et al.*, 2009): participants heard words like *þarricade*, where the initial sound (þ) could be a /b/, a /p/, or anything in between. In critical trials, listeners would initially activate a lexical competitor and then had to revise this decision after hearing the word offset. Kapnoula *et al.* (2021) found that listeners did not differ in their initial commitment to lexical competitors, but when the acoustic distance between stimulus and target was large, gradient listeners were more likely to recover from initial errors. This shows that gradiency allows for higher flexibility at "juggling" lexical representations, thus, facilitating spoken language processing when such flexibility is needed.

One particularly promising direction is the use of the VAS in assessing the role of speech gradiency in the context of bi-/multilingualism and L2 learning. A common assumption is that non-native speech sounds are assimilated to similar-sounding L1 categories and this assimilation hinders the learning of non-native speech contrasts (Best and Tyler, 2007; Kuhl, 1991). Therefore, one may predict that more gradiency in L1 should lead to weaker assimilation of L2 speech sounds, thus, facilitating the learning of a new phonology. In line with this rationale, Kim *et al.* (2020) used the VAS to measure individual differences in phoneme categorization gradiency, which they related to listeners' ability to adapt to unfamiliar vowel categories. The results showed that listeners were able to adapt to the new categories, but gradiency was not a significant predictor of this ability. One possible reason for this null effect is that the critical categories were vowels, which are likely to be processed more gradiently by all listeners (Fry *et al.*, 1962)—that is, maybe there is not enough individual variability in that measure. In contrast, preliminary results by Kapnoula and Samuel (2021), also using the VAS, suggest that gradiency in a consonant contrast (b/p) is positively linked to L2 proficiency.

Developmental work on speech categorization and production is also likely to benefit from the use of the VAS. The main advantage here comes from the fact that children's phonological categories and speech processes are still developing, exacerbating the need to detect subtle, fine-grained differences. For example, Munson *et al.* (2010) asked listeners to rate children's /s/ and /θ/ utterances using a VAS and compared the ratings to transcription data. They found that even when /θ/ and /s/ productions were similarly transcribed as [s], they were rated as significantly different by VAS raters. This result highlights the sensitivity of the VAS task to small differences in children's productions.

Apart from assessing speech in typical populations, the VAS has also proved useful in clinical work. In a recent study, Meyer and Munson (2021) asked more- and less-experienced speech and language pathologists/therapists (SLPs/SLTs) to rate children's utterances using a VAS task. The critical finding was that more experienced speech professionals tended to give fewer intermediate ratings (i.e., they used the scale end points more). This pattern is paradoxical given that gradient ratings reflect richer, more fine-grained information that can be used to guide well-tailored clinical interventions. As a result, this finding highlights the need to better understand what exactly drives this reduced sensitivity and to explore ways in which we can promote the use of continuous rating scales by clinicians at all levels of experience [see also Abur *et al.* (2019), who show that the VAS is a good measure of sentence intelligibility for speakers with Parkinson's disease, as well as Xue *et al.* (2021), who provide extensive support for the VAS as way of assessing intelligibility of pathological speech].

The aforementioned findings speak to the value of the VAS as a tool for basic research on the fundamental mechanisms of speech categorization and related fields like bilingualism, L2 learning, speech development, and communications disorders. In fact, we believe this value to go even beyond the current ways in which the VAS is used. For example, as we mentioned earlier, recent work shows that the critical space in which effects of gradiency appear to be the most robust is on/near category boundaries (Kapnoula and McMurray, 2021). This suggests that we need indices that zoom in on the boundary region to better assess perceptual warping. It is an open question whether current VAS indices are doing that in an optimal way. For example, currently, gradiency is estimated mainly based on the VAS slope, which is calculated from participants' pattern of ratings across the continuum. An alternative approach could be to focus on each participant's boundary and extract an estimate of how much their actual ratings deviate from what would be expected if we assumed a fully gradient function. That is, any deviation from that should, in principle, reflect the degree to which a participant is attracted to the end points (i.e., due to perceptual warping) at their very boundary—where such effects should be at their zenith [see Kapnoula and Samuel (2021) for an example].

One last point is that currently VAS data are most commonly analyzed using logistic-shaped functions; however, this is not an intrinsic property of the task. That is, the VAS is by its very nature flexible in allowing listeners to use it as they wish, and, as a result, the observed rating patterns can have different shapes. This property of the VAS makes it a potentially excellent tool for testing different theories regarding the underlying categorization processes. For instance, we can generate theoretically based assumptions about what the response curve should look like under different theories and assess/compare them using VAS data.

X. CONSIDERATIONS WHEN IMPLEMENTING THE VAS

The basic design of the VAS is straightforward: participants see a scale (typically a line along the center of a screen) with end points marked. A stimulus is presented, and the participants mark on the line where they think that stimulus falls between the end points. However, it is critical to carefully consider the exact implementation to avoid biasing participant responses and to maximize the validity of the measure. First, the way that end points are marked could affect performance. In particular, end points should be obviously identified for the population in question. If using the VAS with young children, text end points might not be clear to all participants. Instead, stimuli could be embedded in words (*beach-peach*), and then images can be used to represent the end points. It may be necessary to explicitly identify the end points to participants before trials begin. There may also be a benefit in varying the sides of the scale used for each end point to avoid biases to certain sides of the display.

Second, the visual scale on the screen should not include obvious anchor points, such as hashes at the end points or in the center. These kinds of markers could affect responses as a form of anchoring (Paul-Dauphin *et al.*, 1999). Ideally, the scale line should have no markers whatsoever, and if possible, even a cursor should be avoided, as its initial location could bias responses. A touchscreen monitor can be used to present the scale without any visual indicators and allow participants to respond easily without the need for preset markers—a marker can be placed after registering a touch response.

Finally, it is essential to explain the task extremely clearly. The VAS is not an ecologically typical task for participants, so it is important to ensure they understand how to use the response scale. Practice trials and trials with non-speech items (e.g., a visual continuum) can help orient the participant to the nature of the task. In particular, care should be taken to use clear and accessible instructions when using the task to compare populations. For example, in a developmental study, it is important that younger participants fully understand the task. Related to this, the task should be explained in exactly the same way across participants, including both verbal/written instructions and gesturing, given that any deviations could bias participants' behavior. Similarly, when using the task cross-linguistically, the instructions should be carefully aligned between languages to ensure that all groups approach the task with a similar understanding.

XI. LIMITATIONS

The VAS overcomes several limitations of previous methods. It separates speech categorization from response mapping and allows estimation of the gradiency of categorization with fewer trials, and individual trials can be meaningfully interpreted. However, the task is not without its own limitations. First, it is possible that participants might adopt a different processing strategy because of the scale—as 2AFC might encourage participants to dichotomize their representations, the VAS might cue them to more closely attend to subcategorical differences. As a result, listeners' general approaches to using continuous rating scales may shape the responses, independent of the actual speech content. Some work suggests that the strongest form of this concern is unlikely; listeners' response functions in the VAS do not usually correlate with their use of the scale for a visual continuum (e.g., images ranging from apple to pear) (Kapnoula *et al.*, 2021; Kapnoula and McMurray, 2021). In addition, it is noteworthy that correlations of VAS slopes within participants are higher for acoustically similar continua (e.g., labial and alveolar stops) compared to dissimilar ones (e.g., labial stops and fricatives; Kapnoula *et al.*, 2017; Kapnoula *et al.*, 2021; Kapnoula and McMurray, 2021). Together, these patterns suggest that listeners do not adopt a particular scaling strategy to use for the task in general. However, the broader concern that they may be cued to attend differently to the perceptual information remains a necessary area for future research.

The VAS might also inadvertently increase memory demands relative to 2AFC, depending on how a listener categorizes stimuli and maps them to response options. If listeners first make category judgments, and thereafter convert these to goodness ratings to perform the VAS task, they might have to rely more heavily on working memory representations of the stimuli. This could lead to increased warping toward category identity in working memory. However, even with such increased warping, the VAS would still be better equipped to assess whether continuous detail persists than in a more rapid response task that forces dichotomous responding. Nevertheless, future research should investigate whether the VAS entails such additional processing stages, perhaps by comparing RTs between the 2AFC and VAS tasks.

Furthermore, much of current research using the VAS has been conducted with English monolingual normal-hearing adults, and the range of phonetic contrasts on which VAS has been used is still very limited. This raises the possibility that the appropriateness of different VAS measures may vary depending on the population and stimuli used. For example, in Spanish, the /b/-/p/ contrast is based on the presence/absence of pre-voicing. Given that this cue may function more like a presence/absence than a continuous difference, this may drive listeners to pay less attention to within-category differences, even when these are perceivable. Consequently, this may lead them to treat the VAS almost like a 2AFC [see also van Alphen and McQueen (2006)]. Indeed, preliminary data from Spanish speakers are in line with this, showing that, in contrast to English speakers, VAS slopes in this population are significantly correlated with 2AFC slopes (Kapnoula and Samuel, 2021).

Interpretation of the VAS relies on estimating parameters of the categorization functions for individual listeners. This process is non-trivial. In particular, although the VAS requires fewer trials than tasks with dichotomous responses, it still may require substantial repetitions to reliably model these parameters. Tools like Bayesian nonlinear models, which simultaneously estimate group- and individual-level estimates, can help (see https://osf.io/q39yt/), but proper statistical analyses for the VAS remain an open area of inquiry.

The VAS also introduces interpretive challenges in the need to consider multiple parameters simultaneously. In particular, slope should be considered in light of response variability to determine whether shallow slopes arise from use of the entire scale or inconsistent use of end points. This combination is not entirely straightforward, however. A steep slope is predicated on low response variability, whereas a shallow slope can have low or high variability. This heteroscedasticity may make it more challenging to use an interaction term to interpret the effect. Instead, other methods, such as latent profile analysis or multivariate regression or analysis of variance (ANOVA), may be appropriate for modeling the two-dimensional space.

Finally, it is important to recognize that the VAS is far removed from ecological speech perception—the task inherently asks participants to make metalinguistic judgments to assess an intermediate stage of language processing. This is necessarily true of any task that asks for sub-lexical judgments. However, such laboratory tasks are often necessary to isolate processes of speech categorization from higher-order processes. In this, the VAS better isolates these tasks from response-mapping processes that are engendered by the 2AFC task. The VAS, thus, offers insight into the nature of processes that serve as one component of the larger language processing system.

XII. CONCLUSIONS

Speech categorization requires a listener to accommodate variability in acoustic realizations of spoken stimuli. However, this plays out across multiple cognitive processes, including auditory encoding, categorization, and response generation, as well as external processes like memory, executive function, and attention. Classic measures of speech categorization conflate these processes under an assumption that identification using discrete choices is a straightforward representation of the full categorization process. We have demonstrated why identification tasks are unable to discriminate between different processes that can produce the same behavioral outcomes. The field needs tasks that are better able to embrace the multifaceted nature of speech categorization. The VAS is a promising candidate for this. This task offers insights into several aspects of speech categorization, including dissociations between encoding, categorization, and response generation.

ACKNOWLEDGMENTS

The authors thank William McEchron for providing methodological inspiration and Jan Edwards for being awesome. This project was supported by National Institutes of Health (NIH) Grant No. DC008089 awarded to B.M. This work was supported by the Basque Government through the Basque Excellence Research Center (BERC) 2018–2021 and BERC 2022–2025 programs, by the Spanish State Research Agency through BCBL Severo Ochoa Excellence Accreditation Nos. SEV-2015-0490 and CEX2020-001010-S, and by the Spanish Ministry of Science and Innovation through Grant No. PID2020-113348GB-I00, awarded to E.C.K. This project has received funding from the European Union's Horizon 2020 research and innovation program, under Marie Skłodowska-Curie Grant Agreement No. 793919, awarded to E.C.K.

[1] See McMurray and Haskins Laboratories (2022) for discussion of why the steepness of identification curves can be misleading.

[2] Which is probably necessary for any experiment.

[3] Because all speech perception tasks need an unintelligible acronym.

[4] See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0015201 for a short video of the task.

[5] There are various ways that these parameters can be estimated and interpreted. Example code to estimate these parameters is available at https://osf.io/4atgv/ (version 31 or later) as a way to help guide analysis and demonstrate our approach. However, other methods may prove more appropriate with further research.

Abur, D., Enos, N. M., and Stepp, C. E. (**2019**). "Visual analog scale ratings and orthographic transcription measures of sentence intelligibility in Parkinson's disease with variable listener exposure," Am. J. Speech Lang. Pathol. **28**(3), 1222–1232.

Allen, J. S., and Miller, J. L. (**1999**). "Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words," J. Acoust. Soc. Am. **106**(4), 2031–2039.

Allen, J. S., and Miller, J. L. (**2004**). "Listener sensitivity to individual talker differences in voice-onset-time," J. Acoust. Soc. Am. **115**(6), 3171–3183.

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (**1998**). "Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models," J. Mem. Lang. **38**(4), 419–439.

Andruski, J. E., Blumstein, S. E., and Burton, M. W. (**1994**). "The effect of subphonetic differences on lexical access," Cognition **52**(3), 163–187.

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., and Yamada, T. (**2004**). "Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/," J. Phon. **32**(2), 233–250.

Best, C., and Tyler, M. (**2007**). "Nonnative and second-language speech perception," in *Language Experience in Second Language Speech Learning: In honor of James Emil Flege*, edited by O.-S. Bohn and M. J. Munro (John Benjamins, Amsterdam, Netherlands), pp. 13–34.

Blomert, L., and Mitterer, H. (**2004**). "The fragile nature of the speech-perception deficit in dyslexia: Natural vs. synthetic speech," Brain Lang. **89**(1), 21–26.

Bosch, L. (**2010**). "The acquisition of language-specific sound categories from a bilingual input," in *Selected Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology*, edited by M. Ortega-Llebaria (Cascadilla Proceedings Project, Somerville, MA), pp. 1–10.

Carney, A. E., Widin, G. P., and Viemeister, N. F. (**1977**). "Noncategorical perception of stop consonants differing in VOT," J. Acoust. Soc. Am. **62**(4), 961–970.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (**2008**). "Perception of speech reflects optimal use of probabilistic speech cues," Cognition **108**(3), 804–809.

Coady, J. A., Evans, J. L., Mainela-Arnold, E., and Kluender, K. R. (**2007**). "Children with specific language impairments perceive speech most categorically when tokens are natural and meaningful," J. Speech Lang. Hear. Res. **50**(1), 41–57.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., and Hogan, E. M. (**2001**). "Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition," Lang. Cogn. Process. **16**(5–6), 507–534.

Dial, H. R., McMurray, B., and Martin, R. C. (**2019**). "Lexical processing depends on sublexical processing: Evidence from the visual world paradigm and aphasia," Atten. Percept. Psychophys. **81**(4), 1047–1064.

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M. (**1962**). "The identification and discrimination of synthetic vowels," Lang. Speech **5**(4), 171–189.

Fuhrmeister, P., and Myers, E. B. (**2021**). "Structural neural correlates of individual differences in categorical perception," Brain Lang. **215**, 104919.

Gerrits, E., and Schouten, B. (**2004**). "Categorical perception depends on the discrimination task," Percept. Psychophys. **66**(3), 363–376.

Getz, L. M., and Toscano, J. C. (**2021**). "The time-course of speech perception revealed by temporally-sensitive neural measures," Wiley Interdiscip. Rev. Cogn. Sci. **12**(2), e1541.

Goriot, C., McQueen, J. M., Unsworth, S., van Hout, R., and Broersma, M. (**2020**). "Perception of English phonetic contrasts by Dutch children: How bilingual are early-English learners?," PLoS One **15**(3), e0229902.

Harnad, S. R. (**1987**). "Category induction and representation," in *Categorical Perception: The Groundwork of Cognition*, edited by S. Harnad (Cambridge University, Cambridge, UK), pp. 535–565.

Hazan, V., and Barrett, S. (**2000**). "The development of phonemic categorization in children aged 6–12," J. Phon. **28**(4), 377–396.

Holt, L. L. (**2006**). "The mean matters: Effects of statistically defined non-speech spectral distributions on speech categorization," J. Acoust. Soc. Am. **120**(5), 2801–2817.

Holt, L. L., and Lotto, A. J. (**2010**). "Speech perception as categorization," Atten. Percept. Psychophys. **72**(5), 1218–1227.

Johnson, K., Strand, E. A., and D'Imperio, M. (**1999**). "Auditory–visual integration of talker gender in vowel perception," J. Phon. **27**(4), 359–384.

Kapnoula, E. C., Edwards, J., and McMurray, B. (**2021**). "Gradient activation of speech categories facilitates listeners' recovery from lexical garden paths, but not perception of speech-in-noise," J. Exp. Psychol. Hum. Percept. Perform. **47**(4), 578–595.

Kapnoula, E. C., and McMurray, B. (**2021**). "Idiosyncratic use of bottom-up and top-down information leads to differences in speech perception flexibility: Converging evidence from ERPs and eye-tracking," Brain Lang. **223**, 105031.

Kapnoula, E. C., and Samuel, A. G. (**2021**). "Does sensitivity to acoustic variation within an L1 phoneme category help L2 learning?," in *Proceedings of the 62nd Annual Meeting of the Psychonomic Society*, November 4–7.

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., and McMurray, B. (**2017**). "Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach," J. Exp. Psychol. Hum. Percept. Perform. **43**(9), 1594–1611.

Kim, D., Clayards, M., and Kong, E. J. (**2020**). "Individual differences in perceptual adaptation to unfamiliar phonetic categories," J. Phon. **81**, 100984.

Kleinschmidt, D. F., and Jaeger, T. F. (**2015**). "Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel," Psychol. Rev. **122**(2), 148–203.

Kluender, K. R. (**1994**). "Speech perception as a tractable problem in cognitive science," in *Handbook of Psycholinguistics*, edited by M. A. Gernsbacher (Academic, New York).

Kocsis, Z., Jenison, R. L., Cope, T. E., Taylor, P. N., Calmus, R. M., McMurray, B., Rhone, A. E., Sarrett, M. E., Kikuchi, Y., Gander, P. E., Berger, J. I., Kovach, C. K., Choi, I., Greenlee, J. D., Kawasaki, H., Griffiths, T. D., Howard, M., III, and Petkov, C. I. (**2022**). "Immediate causal impact and compensation after the loss of a neural hub in the human brain," bioRxiv 2022.04.15.488388.

Kong, E. J., and Edwards, J. (**2016**). "Individual differences in categorical perception of speech: Cue weighting and executive function," J. Phon. **59**, 40–57.

Kraljic, T., and Samuel, A. G. (**2005**). "Perceptual learning for speech: Is there a return to normal?," Cogn. Psychol. **51**(2), 141–178.

Kraljic, T., Samuel, A. G., and Brennan, S. E. (**2008**). "First impressions and last resorts: How listeners adjust to speaker variability," Psychol. Sci. **19**(4), 332–338.

Kuhl, P. K. (**1991**). "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not," Percept. Psychophys. **50**(2), 93–107.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (**1957**). "The discrimination of speech sounds within and across phoneme boundaries," J. Exp. Psychol. **54**(5), 358–368.

Liberman, A. M., and Whalen, D. H. (**2000**). "On the relation of speech to language," Trends Cogn. Sci. **4**(5), 187–196.

Manis, F. R., McBride-Chang, C., Seidenberg, M. S., Keating, P., Doi, L. M., Munson, B., and Petersen, A. (**1997**). "Are speech perception deficits associated with developmental dyslexia?," J. Exp. Child Psychol. **66**(2), 211–235.

Massaro, D. W., and Cohen, M. (**1983**). "Categorical or continuous speech perception: A new test," Speech Commun. **2**, 15–35.

McMurray, B. (**2017**). "Nonlinear curvefitting for psycholinguistic (and other) data," https://osf.io.4atgv (Last viewed September 16, 2022).

McMurray, B., Danelz, A., Rigler, H., and Seedorff, M. (**2018**). "Speech categorization develops slowly through adolescence," Dev. Psychol. **54**(8), 1472–1491.

McMurray, B., and Haskins Laboratories (**2022**). "Reconsidering classic ideas in speech communication," J. Acoust. Soc. Am. **152**, in press.

McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (**2002**). "Gradient effects of within-category phonetic variation on lexical access," Cognition **86**(2), B33–B42.

McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (**2009**). "Within-category VOT affects recovery from lexical garden-paths: Evidence against phoneme-level inhibition," J. Mem. Lang. **60**(1), 65–91.

McMurray, B., Tanenhaus, M. K., Aslin, R. N., and Spivey, M. J. (**2003**). "Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category VOT on lexical access," J. Psycholinguist. Res. **32**(1), 77–97.

McQueen, J. M., Tyler, M. D., and Cutler, A. (**2012**). "Lexical retuning of children's speech perception: Evidence for knowledge about words' component sounds," Lang. Learn. Dev. **8**(4), 317–339.

Meyer, M. K., and Munson, B. (**2021**). "Clinical experience and categorical perception of children's speech," Int. J. Lang. Commun. Dis. **56**, 374–388.

Miller, J. L. (**1997**). "Internal structure of phonetic categories," Lang. Cogn. Process. **12**(5–6), 865–870.

Miller, J. L., and Volaitis, L. E. (**1989**). "Effect of speaking rate on the perceptual structure of a phonetic category," Percept. Psychophys. **46**(6), 505–512.

Munson, B., and Carlson, K. U. (**2016**). "An exploration of methods for rating children's productions of sibilant fricatives," Speech Lang. Hear. **19**(1), 36–45.

Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., and Meyer, M. K. (**2010**). "Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of *Vox Humana*," Clin. Linguist. Phon. **24**(4–5), 245–260.

Nearey, T., and Hogan, J. (**1986**). "Phonological contrast in experimental phonetics: Relating distributions of production data to perceptual categorization curves," in *Experimental Phonology*, edited by J. J. Ohala and J. J. Jaeger (Academic, New York), pp. 141–146.

Noordenbos, M. W., Segers, E., Serniclaes, W., Mitterer, H., and Verhoeven, L. (**2012**). "Allophonic mode of speech perception in Dutch children at risk for dyslexia: A longitudinal study," Res. Dev. Disabil. **33**(5), 1469–1483.

Norris, D., McQueen, J. M., and Cutler, A. (**2000**). "Merging information in speech recognition: Feedback is never necessary," Behav. Brain Sci. **23**(3), 299–325.

Ou, J., and Yu, A. C. L. (**2022**). "Neural correlates of individual differences in speech categorisation: Evidence from subcortical, cortical, and behavioural measures," Lang. Cogn. Neurosci. **37**, 269–284.

Ou, J., Yu, A. C. L., and Xiang, M. (**2021**). "Individual differences in categorization gradience as predicted by online processing of phonetic cues during spoken word recognition: Evidence from eye movements," Cogn. Sci. **45**(3), e12948.

Paul-Dauphin, A., Guillemin, F., Virion, J. M., and Briançon, S. (**1999**). "Bias and precision in visual analogue scales: A randomized controlled trial," Am. J. Epidemiol. **150**(10), 1117–1127.

Perkell, J. S., and Klatt, D. (**2014**). *Invariance and Variability in Speech Processes* (Psychology, London).

Pisoni, D. B., and Tash, J. (**1974**). "Reaction times to comparisons within and across phonetic categories," Percept. Psychophys. **15**(2), 285–290.

Reinisch, E., and Holt, L. L. (**2014**). "Lexically guided phonetic retuning of foreign-accented speech and its generalization," J. Exp. Psychol. Hum. Percept. Perform. **40**(2), 539–555.

Repp, B. H. (**1984**). "Categorical perception: Issues, methods and findings," in *Speech and Language*, edited by N. Lass (Academic, New York), Vol. 10, pp. 244–335.

Roberson, D., Hanley, J. R., and Pak, H. (**2009**). "Thresholds for color discrimination in English and Korean speakers," Cognition **112**(3), 482–487.

Robertson, E., Joanisse, M. F., Desroches, A., and Ng, S. (**2009**). "Categorical speech perception deficits distinguish language and reading impairments in children," Dev. Sci. **12**(5), 753–767.

Salverda, A. P., Dahan, D., and McQueen, J. M. (**2003**). "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension," Cognition **90**(1), 51–89.

Sarrett, M. E., McMurray, B., and Kapnoula, E. C. (**2020**). "Dynamic EEG analysis during language comprehension reveals interactive cascades between perceptual processing and sentential expectations," Brain Lang. **211**, 104875.

Schouten, B., Gerrits, E., and van Hessen, A. (**2003**). "The end of categorical perception as we know it," Speech Commun. **41**(1), 71–80.

Sebastián-Gallés, N. (**2011**). "Bilingual language acquisition: Where does the difference lie?," Hum. Dev. **53**(5), 245–255.

Sebastián-Gallés, N., and Bosch, L. (**2002**). "Building phonotactic knowledge in bilinguals: Role of early exposure," J. Exp. Psychol. Hum. Percept. Perform. **28**(4), 974–989.

Serniclaes, W. (**2006**). "Allophonic perception in developmental dyslexia: Origin, reliability and implications of the categorical perception deficit," Writ. Lang. Lit. **9**(1), 135–152.

Slawinski, E. B., and Fitzgerald, L. K. (**1998**). "Perceptual development of the categorization of the /r-w/ contrast in normal children," J. Phon. **26**(1), 27–43.

Spivey, M. J., Grosjean, M., and Knoblich, G. (**2005**). "Continuous attraction toward phonological competitors," Proc. Natl. Acad. Sci. **102**(29), 10393–10398.

Strand, E. A., and Johnson, K. (**1996**). "Gradient and visual speaker normalization in the perception of fricatives," in *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference: Bielefeld, October 1996*, edited by D. Gibbon (Mouton de Gruyter, Berlin), pp. 14–26.

Sumner, M. (**2011**). "The role of variation in the perception of accented speech," Cognition **119**(1), 131–136.

Sussman, J. (**1993**). "Perception of formant transition cues to place of articulation in children with language impairments," J. Speech Lang. Hear. Res. **36**(6), 1286–1299.

Theodore, R. M., Monto, N. R., and Graham, S. (**2020**). "Individual differences in distributional learning for speech: What's ideal for ideal observers?," J. Speech. Lang. Hear. Res. **63**, 1–13.

Toscano, J. C., and McMurray, B. (**2010**). "Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics," Cogn. Sci. **34**(3), 434–464.

Toscano, J. C., McMurray, B., Dennhardt, J., and Luck, S. J. (**2010**). "Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech," Psychol. Sci. **21**(10), 1532–1540.

Utman, J. A., Blumstein, S. E., and Burton, M. W. (**2000**). "Effects of subphonetic and syllable structure variation on word recognition," Percept. Psychophys. **62**(6), 1297–1311.

van Alphen, P. M., and McQueen, J. M. (**2006**). "The effect of voice onset time differences on lexical access in Dutch," J. Exp. Psychol. Hum. Percept. Perform. **32**(1), 178–196.

Vane, J. R., and Motta, R. W. (**1980**). "Test response inconsistency in young children," J. School Psychol. **18**(1), 25–33.

Werker, J. F., and Tees, R. C. (**1987**). "Speech perception in severely disabled and average reading children," Can. J. Psychol. **41**(1), 48–61.

Wichmann, F. A., and Hill, N. J. (**2001**). "The psychometric function: I. Fitting, sampling, and goodness of fit," Percept. Psychophys. **63**(8), 1293–1313.

Xue, W., van Hout, R., Cucchiarini, C., and Strik, H. (**2021**). "Assessing speech intelligibility of pathological speech: Test types, ratings and transcription measures," Clin. Linguist. Phonetics (published online).