

# Evidence and methods in phonology: Supplement material

Paul de Lacy  
Rutgers The State University of New Jersey  
2025

Citation:  
de Lacy, Paul (2025). *Evidence and methods in phonology: Supplementary material*. ms., Rutgers University. <https://www.pauldelacy.net>

The following paragraphs were removed from the chapter “Evidence and methods in phonology” in order to stay within the word limit for the *Handbook of Phonology*. They are provided here as supplementary text.

## Contents

|                                          |    |
|------------------------------------------|----|
| 1. Williams (1917) and L1.....           | 1  |
| 2. Multiple Sources.....                 | 2  |
| 3. Lexicons.....                         | 4  |
| 4. Style-Shifting.....                   | 7  |
| 5. Current Evidence.....                 | 7  |
| 6. Fieldwork certainty.....              | 7  |
| 7. Forced Aligners.....                  | 8  |
| 8. Grammaticality and Acceptability..... | 8  |
| 9. Extended Conclusion.....              | 8  |
| END.....                                 | 10 |

## 1. Williams (1917) and L1

[In 25.2.2, after “However, it could have been the other way around – I don’t know.”]

By the way, in terms of L1 status, Williams (1917) illustrates an interesting danger with dictionaries. Did Williams obtain his data from L1 speakers? While he says that the “most natural procedure” was to ask an “intelligent Maori of the older generation”, he then continues: “occasionally a Maori may, in all honesty, have been habitually misusing a word in his own tongue. Moreover, local usage in respect of words is frequently very curious and perplexing.” These comments suggest that Williams did not hold native speaker utterances in high regard. In fact, his earlier comments indicate that he preferred written sources compiled by non-native speakers (e.g. Mr. Colenso, the Reverend Father Becker, and so on). Williams’ comments emphasize that dictionaries and grammatical descriptions are not written with cognitive scientists in mind. Consequently, explicit arguments must be made that the data from dictionaries and grammars actually meet the criteria that phonologists need for their theories – i.e. that dictionary and grammar data is sourced from a P\*.

## 2. Multiple Sources

[In 25.2.1, just after “So, there is no guarantee that there is a single PhM state that is responsible for the data in (1), so the data in (1) cannot be evidence for that PhM state.”]

Why do so many researchers merge data from *multiple* individuals and present that as evidence that there is a *single* PhM state that can generate that data? I think there are many reasons: tradition, an assumption that speakers of the same language have identical PhM states, and a belief that results from multiple speakers will drown out any impaired or imperfect speakers. None of these turn out to be good reasons.

Regarding tradition, the study of ‘phonology’ is much older than cognitive science, and the object of study used to be ‘languages’. The idea that phonologists research languages and not PhM states continues to persist. In fact, if you look back at the text just before (1) [in the *Handbook* chapter] and the caption for (1), you will notice that I subtly did *not* say that the data was evidence for a PhM state, but from “the Polynesian language Māori”. The idea that “language” and “PhM state” are essentially the same is a profound problem for evidence. A ‘language’ is the shared (or perceived-as-shared) communicative system between multiple people. In contrast, a PhM state exists in an individual. So, descriptions of languages are not necessarily descriptions of PhM states.

Another reason may be the idea that if you get multiple people with the *same* PhM state, you can legitimately pool their data. If A, B, C say different words but have the same PhM state, then *if* A, B, and C had *each* said *all* of the words, they would have said them all the same.

The problem is how we *know* that A, B, and C have the same PhM states. The only way is to compare A, B, and C’s outputs. However, if we have to do all that, why not just elicit all of the data from A?

If we compare A, B, and C’s outputs for one phenomenon and find that they’re the same, can we just assume that A, B, and C’s outputs for another phenomenon will all be the same? (e.g. suppose that A, B, and C all assimilate the same; can we assume that they have the same stress pattern?). Unfortunately, I don’t know of any reason to make this assumption: this belief assumes a level of knowledge about the interdependence of parts of PhM states that we do not have.

I worry that the reason that data pooling is so common is because of informal proxies for determining identical PhM state, such as “If A, B, and C all speak the same language, they must have the same PhM state.” Echeverría (1964) explored this issue in an interesting way: by asking each speaker if they thought the others were speaking the same language as them. They agreed that one of them was not, showing that even the assumption that two people who say they speak the same language are mutually intelligible is misguided.

The belief in PhM state identity also assumes that there is very little inter-speaker phonological variation. In my fieldwork and experiments, I have found that phonological systems show surprising variation. For example, In early 2025, I traveled around New Zealand eliciting flapping and l-vocalization. For a supposedly homogenous mono-dialectal country, there is a remarkable amount of variation in whether people flap or l-vocalize and where they do it (see de Lacy 2026).

The additional challenge is that our evidence is not generated by PhM states alone. The data in (1) [of the *Handbook* chapter], for example, reports observations that – if from an individual P\* – are generated by four cognitive modules (Lexicon, Morphology, Phonology, Phonetics), a neuro-motor interface, and speech articulators. So, if we pool data from two individuals, we are not only saying that their PhM states are the same, but that their Lexicon and Morphology modules are also the same (in whatever way relevant to the data), and that their downstream modules and mechanisms are also the same in whatever way is evidentially crucial. I suspect it would be practically impossible to find two speakers who have identical Lexicons and PhM states, let alone any other module.

Another reason to have multiple speakers is that if there is something wrong with one speaker, the others will ‘drown out’ that speaker’s misleading influence on the results. For example, if one speaker is intoxicated, then the other non-intoxicated speakers’ results will pull the mean measurement towards an accurate value. The problem, of course, is that the majority of speakers might be intoxicated, so we’d have to test every speaker for intoxication anyway. More generally, the idea that the majority of subjects will drown out effects of impairment, atrophy, non-L1ness, non-dominance, and so on has not been demonstrated. In my experience, at least for certain demographics, deviation from P\* is the norm. For example, I suspect undergraduate university students will rarely use their dominant L1 in laboratory experiments, instead defaulting to a formal speech style (which is either an L2 or non-dominant L1). Adding more students will only drown out the dominant L1 speakers more.

I can only think of one reason to study human speech using multiple speakers: if the object of study is not the PhM, but a concept such as ‘language’. A ‘language’ is a communicative system that is shared between multiple people, so a description of a language should take the common elements of speakers of that language, so the more speakers the better (depending on exactly which language you want). In fact, a ‘language’ is a concept that exists independently of its speakers, that’s why dictionaries are usually the amalgamation of many speakers’ vocabulary across speech styles, dialects, and time. In contrast, cognitive scientists are interested in individual speaker’s Competence: a cogsci ‘dictionary’ is a description of a speaker’s Lexicon.

Finally, a justification for pooling multiple individuals’ results may come from assumptions about learning. As Jo (2024:137) puts it, some indicate a belief that “individuals who received roughly the same linguistic input converge on the same linguistic knowledge (grammar), and that individual differences arise due to random variation or differences in factors related to performance or language use”. However, what does ‘roughly the same linguistic input’ mean? And why think that even the *same* linguistic input would lead to the same grammar in every individual? And what about individual phonological innovation? Why not assume that everyone receives profoundly *different* input, varying not only frequencies of segment, word, and sentence type, but also in order of presentation? It seems to me that until we know exactly which input will lead to the same P\* in two individuals, we should assume that it does not, and so pooling data across individuals (let alone across PhM states) is not justified.

After all, when individual results are given in phonological research, there are often not only phonetic differences, but phonological ones, even by speakers of the ‘same’ language. For example, Gordon (2004)’s description of Chickasaw word stress showed that speakers varied as to which phonological distinctions they made in word stress (primary vs. secondary vs. unstress or primary vs unstress, or primary and secondary vs. unstress). Braver (2014) showed that American English speakers distinguished /t/-derived flaps from /d/-derived flaps, while others did not. Shih (2026) shows that several European Portuguese speakers have profoundly different vowel reduction systems.

Bennett & Braver (2020) address this issue head-on for Xhosa labial palatalization. They conclude that “different speakers can have fundamentally different grammatical representations of the same sound pattern.”

In my own fieldwork, I have never found two speakers of the same ‘language’ that have the same Lexicons or phonological systems. For example, B knew words that H did not, and B had a morpheme (the past passive) that H did not.

Relevantly, Kingston et al. (2015) observe that “Listeners’ responses in routine speech perception experiments often differ considerably from one another.” (and conclude that “the search for explanation has just begun”).

Jo (2024) is an extensive examination of individual differences for a phonological phenomenon, and aims to determine the sources of those differences. While certain variation in properties of the observations can be ascribed to extra-phonological factors, Jo (2024:146) concludes “inter-speaker variation ... is at least partially rooted in differences in grammar”.

So, the idea that individual differences among certain populations (however defined) are due to Performance (i.e. PhM-external) and not Competence (i.e. PhM-internal) factors is something that must be proven rather than assumed.

I think the feeling that pooling multiple peoples’ results is legitimate is fueled by the fact that individual results in phonological research are rarely presented. It is also fueled explicitly in some cases by journals that require multiple speakers and pooled results. If individual results *were* broken out more often, I think we would start to see that pooled data is something that is rarely appropriate for use in cognitive science in practice, not just in theory.

Of course, I would go further and say that it is also a mistake to pool data from different *states* within the same individual. Put simply, the proper number of subjects in any phonology experiment or fieldwork is not one, but *less than one* – i.e. a single module state of the many that exist in each individual.

Let’s turn the question around: is it *ever* legitimate to assume that two speakers have the same PhM state? Theoretically. Suppose that two speakers are exposed to exactly the same speech stimuli throughout their lives (e.g. conjoined twins). They might then have the same PhM states. Or perhaps not – the assumption is that for any set of stimuli there is one PhM state that the learner will converge on. Other than conjoined twins, though, speakers will receive different stimuli in different orders, so assuming that any two people have the same PhM state, or even the same *part* of a PhM state, requires explicit, extensive justification.

Bennett, Will & Aaron Braver (2020). Different speakers, different grammars: Productivity and representation of Xhosa labial palatalization. *Phonological Data and Analysis* 2.6: 1-29.

James, Ariel N., Scott H. Fraundorf, Eun-Kyung Lee & Duane G. Watson (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language* 102. 155–181. doi:10.1016/j.jml.2018.05.006

Kingston, J., Rich, S., Shen, A., and Sered, S. (2015). Is perception personal? *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK. 5pp.(<http://www.icphs2015.info/pdfs/Papers/ICPHS0887.pdf>)

### 3. Lexicons

[In 25.2.1, after “In other words, creative combination of data from multiple speakers can create a dataset that is not generated by *any* P\* (so-called ‘Frankenstein PhMs’; [de Lacy 2009](#)).”]

To expand on this point, the danger of creating a Frankenstein PhM when using a dictionary, grammar, or corpus as a source cannot be overstated. It’s usual practice for such sources to combine data from multiple speakers – in fact, it’s *expected* and dictionaries/grammars/corpus based on a single source are usually seen as inadequate. So, if you see a dataset and its source is a dictionary/grammar/corpus, you can very reasonably assume that the data was combined from multiple PhMs. Even though presentation of the dataset implies that there is *some* individual PhM that generates all of cited the data, it is very rare to find clear evidence that this is so.

I’ll make this point more elaborately by building on the Māori data above. I presented three passive forms [arumia], [motokia], and [waruhia]. However, as mentioned above, Williams’ dictionary does not claim that the same person produced all of these forms, let alone the same PhM.

Let's add two doses of reality. The first is that if you talk to Māori native speakers, you'll find individuals who know one or two of these words, but not all of them. Taking the possibility of not knowing a word into account, there are eight potentially attested output lexicons, given below. I use 'Ø' to mean 'doesn't know the relevant form'. For example, (a) is an individual who knows *arumia*, but not *motokia* or *waruhia*.

(X) *Possible output lexicons*

- (a) [arumia], Ø, Ø
- (b) Ø, [motokia], Ø
- (c) Ø, Ø, [waruhia]
- (d) [arumia], [motokia], Ø
- (e) [arumia], Ø, [waruhia]
- (f) Ø, [motokia], [waruhia]
- (g) [arumia], [motokia], [waruhia]
- (h) Ø, Ø, Ø

Of these eight output lexicons, which do we actually find among Māori native speakers? Based on Williams' dictionary alone, we don't know. What we do know is that there was some individual who produced *arumia*, some person who produced *motokia*, and someone who produced *waruhia*. We don't know whether it was the same person.

What we do know is that how ever many different people there were, their lexicons combine to produce all of *arumia*, *motokia*, and *waruhia*. Unfortunately, that could mean that there were just three types of people: (a), (b), and (c). None of these individuals provide evidence for a phonological system with final consonant deletion.

It's also possible that there were two types of individual: (a) and (f). In terms of evidence for final consonant deletion, (f) is on the right track, but just two forms aren't convincing.

To add some more reality, dictionaries provide *alternative* passive forms for the three words. For *aru*, two possible passives are listed: *aruhia* and *arumia*. For *moto*: *motokia* and *motoa*, and for *waru*, *waruhia* and *warua*.

Let's assume that everybody has just one of these forms (i.e. no free variation). We're left now with 27 possible lexicons (i.e. {aruhia, arumia, Ø} × {motokia, motoa, Ø} × {waruhia, warua, Ø}). Which of these lexicons actually exist? I don't know, and there's no way to tell from the dictionaries alone. As above, at minimum we'd need three to exist, but only some of them provide evidence for final consonant deletion. More importantly, we cannot know for any particular lexicon whether there is a person who has that specific lexicon.

To make matters more realistic and even more complicated, some individuals could have their passive forms in free variation – a person might say *aruhia* some of the time and *arumia* at other times: i.e. {aruhia, arumia, aruhia~arumia, Ø} × {motokia, motoa, motokia~motoa, Ø} × {waruhia, warua, waruhia~warua, Ø} = 64 possible lexicons. Given the nature of our data source, we do not know which of these 64 possible lexicons are actually attested. Some of them are, but we don't know which ones, and we don't know whether they support the claim that there is word-final consonant deletion.

Finally, we need to ask about the roots *aru*, *moto*, and *waru*. If an individual has the passive form but no root, is the passive form suppletive for that speaker? For example, if a speaker has *arumia* but no *aru*, could the speaker learn that the underlying form of the root is /arum/? Let's suppose the answer is *no*. However, we know that it is possible for speakers to have derived forms without having their roots (e.g. I can say inherently plural word *pajamas*, but I don't have the root *pajama* as an independent word). So, we need to independently vary the possible lexicons for the

roots. In other words, it's possible that there's a speaker with *arumia*, *motokia*, and *waruhia*, but no *aru*, *moto*, and *waru*.

We can now describe our possible lexicons as:  $\{\text{aru}, \emptyset\} \times \{\text{moto}, \emptyset\} \times \{\text{waru}, \emptyset\} \times \{\text{aruhia}, \text{arumia}, \text{arumia} \sim \text{arumia}, \emptyset\} \times \{\text{motokia}, \text{motoa}, \text{motokia} \sim \text{motoa}, \emptyset\} \times \{\text{waruhia}, \text{warua}, \text{waruhia} \sim \text{warua}, \emptyset\}$ . So, there are 512 possible lexicons.

Let's now think about which of these lexicons provide evidence that underlying final consonants delete. Both *aruhia* and *arumia* can be used as evidence that *aru* underlyingly ends in a consonant, supposing that some speakers have /aruh/ and others /arum/. However, *motokia* and *motoa* are different: for speakers with *moto-a*, it is clear that *moto* does not end in a consonant underlyingly. The same goes for the form *waru-a*. Of the passive forms, only three can be used to argue that the roots underlyingly end in a consonant: *aruhia*, *arumia*, *motokia*, and *waruhia*.

Suppose the learner hears *aru*~*aruhia* but *moto*~*motoa* and *waru*~*warua*. Would you the learner conclude that *aru* was underlyingly /aruh/? Possibly not – the learner might conclude that *aruhia* is lexically listed. So, let's suppose that the learner has to hear at least three forms to break out of the lexical listing hypothesis.

At this point, we're able to answer a question about the data in (1): what are the chances that there is an individual who has a lexicon that would lead them to generate final consonant deletion? Such a lexicon requires both *waru* and *waruhia*, both *moto* and *motokia*, and *aru* with either *aruhia* or *arumia*. If a person lacks any of these pairs, we assume that they decide the forms are suppletive.

In other words, of the 512 possible lexicons, there are – we assume – only two that would lead a learner to conclude that there is final consonant deletion:  $\{\text{aru}, \text{aruhia}, \text{moto}, \text{motokia}, \text{waru}, \text{waruhia}\}$  and  $\{\text{aru}, \text{arumia}, \text{moto}, \text{motokia}, \text{waru}, \text{waruhia}\}$ . So, when I present the data above, I am asking you to believe that it was elicited from speakers with one of these two lexicons, and not from a group of speakers with the other 510 lexicon types.

But let us go further. Since we *do* have the data *aru*, *arumia*, *moto*, *motokia*, *waru* and *waruhia*, what are the chances that this data did *not* come from an individual who had all of these words in their lexicon? To put it slightly differently, if the data came from two or more individuals, how many unions of lexicons produce  $\{\text{aru}, \text{arumia}, \text{moto}, \text{motokia}, \text{waru}\}$  where none of the joined lexicons are themselves  $\{\text{aru}, \text{arumia}, \text{moto}, \text{motokia}, \text{waru}\}$ ?

After all, it's possible that six different lexicons were amalgamated: one had *aru*, one had *arumia*, another *moto*, and so on. It's also possible that five different lexicons were involved: one had *aru* and *moto*, another had *aruhia* and *waru*, and so on. Neither of these groups have any individual who actually has all six words, and therefore none has evidence for final consonant deletion.

The answer is that there are 202 ways to combine lexicons to produce *aru*, *arumia*, *moto*, *motokia*, *waru* and *waruhia* where none of the lexicons has all of the words. So, when we look at the data, we ask "What are the chances that there was a speaker with a lexicon with all of these words?" If we assume that lexicons vary equally, the answer is a little less than half a percent (0.493%).

Of course, some output lexicons might be more likely to exist than others. For example, it might be rare for anyone to know all three words, and equally as rare for anyone to not know any of the words. If so, the likelihoods can change markedly.

So, what should you do when you see a set of data? It's clearly important to ask "Is this data generated by an individual PhM?" If the source mixes the outputs of multiple PhMs (i.e. it has words or sentences from different people or speech styles), then it cannot be assumed that it was generated by an individual PhM. Of course, making such an assumption is not consistent with the cautious stance that is standard operating procedure in science – i.e. don't believe without evidence. If we are sure of our uncertainty, then we must ask whether that uncertainty invalidates the theoretical and/or evidentiary claim being made.

## 4. Style-Shifting

[In 25.2.2, after “I did not ensure that H and B used the same PhM state for every word.”]

I have been in many situations in fieldwork and experiments where it has become obvious that subjects were switching between multiple PhM states (dialects and speech styles). A few years ago, Patrik Bye and I were working with T, a speaker of Nevisian Creole. T pronounced *person* as [peɪsən], but as he got more tired and hungry, he started pronouncing it as [pesən]. Finally, just before lunch, T pronounced *person* as [posən]. Which of these was generated using his L1 Lexicon and P\*? From other descriptions, I suspect it was [posən] (Cooper 1979). However, the incident emphasizes that it is unreasonable to start eliciting data from someone and expect that they will stick to one PhM state, and that this state will be P\*. (By the way, I’m not suggesting that hunger is a good way to induce a P\*; hunger is an impairment, and so could interfere with evidence-elicitation).

## 5. Current Evidence

[In 25.5]

What about current evidence for phonological theories? The vast majority of current evidence is of undetermined value (de Lacy 2012): I don’t know if it’s bad or good; I just don’t know. I’m not alone in this worry. A phonologist who spent years writing about word stress recently told me he’d given up because he couldn’t trust the evidence. Similarly, Zhang (2010:1141), in discussing research on tone in Chinese languages, comments that “uncertainty of impressionistic transcriptions” means that studies must “rebuild an empirical foundation from which theoretical analyses may proceed”.

How serious is the current situation? Over the past 30 years, I have learned methods for acquiring phonological evidence, applied them, and discovered that they produced unreliable and inaccurate results. Consequently, I have no confidence in any of my theories.

For example, I spent many years working on sonority-driven stress using the IEM and grammar-based methods of evidence acquisition, only to have Shih (2018a,b) show that my results were essentially hallucinations obtained through unreliable and inaccurate methods. My students and I have also looked at tone-driven stress and default-to-opposite stress, and found the sources to be unreliable (de Lacy et al. to appear). In fact, in trying to revise my work on markedness, I have found myself despairing at the unreliability of evidence for epenthesis, neutralization, assimilation, and vowel reduction that I relied on in other work (de Lacy 2006). Without any evidence to guide my theorizing, I am afraid I am not doing science, but mathematics.

de Lacy, Paul (2006). *Markedness: Reduction and Preservation in Phonology*. Cambridge University Press.

de Lacy, Paul (to appear). *Non-metrical stress from the Rutgers Phonology Lab*.

## 6. Fieldwork certainty

[End of 25.3.2]

If the IEM is not reliable, why has it formed the basis (via grammars and dictionaries) of so much of the evidence for phonological theories? I believe part of the reason is that the style and presentation of grammars and dictionaries force fieldworkers to express IEM results with a certainty that they rarely feel. Cheliah & de Reuse (2011:256) [CR] comment that “A common experience reported by fieldworkers is a tentativeness or skepticism about their transcriptions in the early stages of

fieldwork.” In my experience, uncertainty isn’t just a feature of the *early* stages of fieldwork, and CR seem to also agree: “There will always be a set of sounds which the fieldworker will have persistent doubt”. However, there is no mechanism for recording the many facets of uncertainty in dictionary entries or grammars, so what ends up in print implies a lot more certainty than what there is.

CR also say that “Many times these early transcriptions turn out to be accurate, and later “corrected transcriptions” turn out to be wrong!”. CR’s comment is essentially the definition of unreliability: where the same person listening to the same recording changes their mind over and over again. To be honest, for me fieldwork using the IEM is a roller-coaster of self-doubt that pauses briefly when I’m forced to write down a transcription, but never really ends.

## 7. Forced Aligners

[In 25.5]

There are also relevant studies on using ‘forced aligners’ – i.e. automated speech segmentation software. MacKenzie & Turton (2020) conclude that “forced alignment software is highly reliable and improving continuously” and is “generally highly accurate for a wide range of varieties.” Of course, exactly how reliable and accurate depends on context. For example, Goldman (2011) found for spontaneous English speech, that the software EasyAlign placed around 50% of segment boundaries within 10ms of the two human raters’ placement. This agreement rate is probably still too low for use in a lot of phonetic and phonological research (although the two humans only agreed at a 62% rate). However, the point is that we *know* about how reliable and accurate forced aligners are, so we can decide whether they are appropriate to use, and if we use them how to correct the values they produce.

## 8. Grammaticality and Acceptability

[In 25.4.2, after “The only reason to use multiple methods is if it was demonstrated that using those methods together increases reliability, accuracy, or precision, compared to using a single method.”]

Has it been proven that the methods used to elicit grammaticality judgements are good (e.g. Schütze 1996, Sprouse et al. 2013 [S13])? For example, during fieldwork, I have often asked questions like “Is [bnɪk] a possible word for you?”, or something similar. (There are other methods, too, such as rating tasks, and forced-choice tasks, and introspection.) There has been a great deal of work into grammaticality and acceptability judgements in syntax (e.g. S13 and references cited therein). S13 showed that the introspection method and several other acceptability tasks correlated well at 95% (margin of error of 5.3-5.8%). This correlation indicates that the methods can produce very similar results, and might be reliable. However, are they valid and accurate?

The goal of such ‘acceptability’ tasks is to determine grammaticality. So, how accurate are acceptability tasks in identifying grammatical and ungrammatical structures? A difficulty is in determining how acceptability and grammaticality are connected – what role does grammaticality have in determining acceptability? Without knowing whether acceptability tasks are construct-valid (i.e. how they reflect grammaticality), it is difficult to know what status to assign them. Some research on linguistic methods use ‘reliability’ and ‘accuracy’ interchangeably, and so some care has to be taken in their interpretation.

## 9. Extended Conclusion

[In 25.5]

Having pervasively uncertain evidence has serious consequences. We may have written off entire classes of theory because we believed they failed to account for evidence that is actually unreliable and inaccurate. Perhaps our field's theoretical volatility is due to the lack of good evidence. And with so much uncertainty about our evidence, what is really determining the 'success' of theories?

I think the situation is so dire that we phonologists should drop everything we're doing and focus on methods. Here's what I think needs to happen field-wide.

We – or at least most of us – need to agree that studying methods is as important as studying theories and producing fieldwork and experimental results. If this belief takes hold, I think changes will follow naturally. Phonology courses will address how evidence is obtained and how to evaluate methods. Students will be encouraged to write MA and PhD theses about method reliability, accuracy, and precision (e.g. Fremed 2013, Schütze 1996). Jobs will go to people who write their PhDs about methodology. Journals will accept articles about methods, and reviewers will require authors to discuss the validity, reliability, accuracy, and precision of their methods.

The challenge is to do all of this without losing what's good about phonology: an emphasis on thoroughly exploring rigorously defined theories. After all, there's no value in being great at gathering evidence for theories if we have no theories to find evidence for.

We are fortunate because the study of methodology is very strong in a number of allied and relevant fields. I have given examples above for measuring onset of voicing and alcohol intoxication. When it comes to speech impairments, there is a lot of research on valid methods (e.g. Bernthal 2009, Akipinar et al. 2010). There is also a strong literature on perception of auditory and speech contrasts that we can build on to determine fieldwork method validity. Professional audio recording equipment always has information on accuracy and precision.

However, we also need to identify where our knowledge is lacking. I believe don't yet know about the validity, reliability, accuracy and precision of methods used for determining whether experiments or fieldwork are eliciting from P\* vs. some other PhM type. We also don't know enough about the value of methods used for finding many crucial acoustic landmarks used in phonological research. We do not know about the value of methods involving using grammars or dictionaries as sources of data, of tasks-specific versions of the IEM.

I am not saying that we can rely on researchers in other fields to look at methods for us. While we can build on their work (e.g. Ray 2009), we cannot assume that people in other fields have the same concerns as phonologists. For example, a lot of research on methods for testing alcohol intoxication is oriented towards determining whether people are too drunk to drive. In contrast, we are worried about whether people are too drunk to elicit from, to perceive stimuli, or make grammaticality judgements. It's possible that the Blood Alcohol level for phonological purposes is different from the one needed for safe driving. So, it's possible (even likely) that the SFST is not appropriate for use in phonological research. However, the work on the SFST is an invaluable starting point for us.

We similarly cannot rely on phoneticians, even for methods involving acoustic or articulatory measures. Phoneticians don't necessarily have the same theoretical goals as phonologists. On the other hand, that doesn't mean not working together (see below).

We also cannot hope that the field's movement towards laboratory experiments (or any other type of evidence source) is going to improve our understanding of methods. We will only advance if we specifically study method reliability, accuracy, and precision, regardless of which evidence source we use. Humans may be better than machines for some tasks (e.g. determining semantic content of morphemes). Also, methods overlap significantly, whether our evidence is sourced from grammars, dictionaries, corpuses, experiments, or fieldwork. In all cases, we must ask about whether

our source is from a P\*, and the methods we use to evaluate whether the source will be similar across evidentiary modalities.

To be clear, not all phonologists should be doing lab experiments. It would be a huge mistake at this stage for everyone to only acquire evidence through lab experiments because doing so incorrectly assumes that we know that all other methods are less reliable, accurate, precise, and practical. No matter how evidence is acquired, what matters is to care about the validity, reliability, accuracy, and precision of the methods used.

I also believe that the way forward is to extend the goals of the Laboratory Phonology movement. I believe that LabPhon has been the most significant movement in advancing phonology over the past few decades. It aimed to get phonologists and phoneticians to work together, and “to promote more integrated methodologies” (Cohn 2010). However, we must go further. The PhM is a very embedded module; there is no device that can detect its outputs or inputs directly. Upstream modules like the Lexicon, Morphology, and Syntax are the source of its inputs, and its outputs pass through the Phonetic Module, the neuro-motor interface, articulators, transmission medium, and every other part of the evidence acquisition process before we can observe the PhM’s effects. The only way forward is to understand every part of this process. So, groups of specialists in each module and mechanism in the evidentiary process must work together in determining exactly what PhM theories – and theories of every other module and mechanism – predict in terms of observable effects. The alternative is for phonologists to learn everything they need about syntax, morphology, articulatory phonetics, acoustic phonetics, perception, first and second language acquisition, alcohol intoxication, aphasia, every other relevant impairment, and statistics. Sadly, all of these fields are too complex and specialized for one person to master them all. In short, wide-ranging cross-field collaboration is essential. Good evidence can only be acquired in teams.

Should theoreticians pause theorizing? No. Without well-defined theories, there is nothing to test. Personally, I can’t resist writing about phonological theories *and* presenting evidence for them; I am just far more circumspect about what I’m actually doing (e.g. Staroverov & de Lacy this volume). My hope is, though, that within the next few decades, someone will be able to figure out how good the evidence is for (or against) our theories.

### **Additional References**

Akpınar, Meltem Esen, İsmail Kocak, Berk Gürkınar, and Halil Ertürk Esen 2010. Effects of soft palate implants on acoustic characteristics of voice and articulation. *Journal of Voice* 25.3: 381-386.

END