Moving past indirect proxies for language experience: 'Native speaker' and residential history are poor predictors of language behavior

Lauretta S. P. Cheng[1] (lspcheng@umich.edu)
Mathew A. Kramer[1] (arkram@umich.edu),
Ria Upreti[2] (upreti.ria@gmail.com),
Savithry Namboodiripad[1] (savithry@umich.edu)

[1]Department of Linguistics, 440 Lorch Hall, 611 Tappan Avenue, Ann Arbor, MI 48109 USA
[2]Department of Linguistics, 305 E. 23rd Street, Austin, TX 78712 USA

Abstract

As widely acknowledged in the bilingualism literature, language experience is multifaceted, complex, and dynamic; it cannot be simply reduced to single dimensions or categories. However, cognitive science research outside of bi/multilingualism does not always take into account this fact. Within a population of Hindi-Urdu speakers, we show that proxy categories based on 'native speaker' identification or residential history do not neatly map onto patterns of language experience, despite the common assumption that these bring about sufficient homogeneity. Moreover, compared to variables derived from gradient measures of language experience, these proxies do not robustly predict linguistic behavior in the form of acceptability judgments in Hindi-Urdu. In demonstrating alternative approaches to operationalizing language experience, we argue for all language researchers to move past relying on underspecified and ideologically-linked concepts, in favor of more intentional, nuanced, and rigorous testing of experiential factors underlying language processing.

Keywords: language experience; native speaker; clustering analysis; acceptability judgment experiments; multilingualism

Motivation

The role of language experience is central to many research questions in cognitive science. However, LANGUAGE EXPERIENCE is a multifaceted construct that is challenging to measure (Marian & Hayakawa, 2021; Gullifer et al., 2021). Though known to be dynamic across the lifespan and heterogeneous within communities, the way that it is operationalized often does not reflect this, as discussed by bilingualism researchers (Luk & Bialystok, 2013; Tsehaye, Pashkova, Tracy, & Allen, 2021). To demonstrate to a broader audience both why and how ideologically laden categorical proxies for language experience should be avoided, we compare the use of 'native speaker' identification and residential history with more direct operationalizations of language experience.

The **native speaker** construct can be used unreflectively in the cognitive sciences—particularly in research focusing on 'monolingual' language processing or knowledge—with the assumption that (a) everyone has a similar understanding of what it means, and (b) it can be used to construct participant groups which are sufficiently homogeneous for experimental work. In opposition, Cheng et al. (2021) detail a variety of methodological and ethical reasons that psycholinguists may wish to avoid using this construct (see also Rampton, 1990; Dewaele, 2018; Dewaele, Bak, & Ortega, 2021), including how 'native speaker' is interpreted differently across contexts by researchers *and* participants (Faez, 2011). These works and others (e.g. Debenport, 2011) show that 'native speaker' represents a language ideology and is often understood by participants as an identity label. Although related, identity is not directly derived from amount or type of linguistic experience (cf. language *allegiance* vs. *expertise*; Rampton, 1990). As such, we do not expect categories based on 'native speaker' to correspond straightforwardly or uniformly to measures of language exposure and use.

Another proxy for language experience is **residential history**, particularly when researchers study language use in diaspora communities. Recent research continues to compare the behavior of *in situ* ("homeland") populations with that of *ex situ* ("heritage") individuals (e.g. y Cabo, 2020; Uygun, Schwarz, & Clahsen, 2021; Kim, 2020; Fernández-Dobao & Herschensohn, 2021; see Ortega, 2020 for critiques). Such an approach assumes that individuals growing up in different environments will have consistently different profiles of experience. While community-level factors such as language policy can indeed have consistent effects on individuals' patterns of language exposure and use, these assumptions may not hold up to scrutiny where individual situations vary widely (e.g. Unsworth, 2019). Like 'native speaker' categorization, we may expect there to be an association between residential patterns and overall language experience, but not necessarily a strong or direct one.

Besides holding only indirect links to language experience, an additional issue with 'native speaker' and residential history are that they are *categories*. As language experience is a multi-factorial construct, recruiting participants and constructing analyses based on a simple categorical proxy can seem an attractive way to handle this inherent multidimensionality. In bilingualism research, however, scholars have shown that continuous measures capture participants' language experience better than categorical labels such as "monolingual" or "bilingual" (Luk & Bialystok, 2013; Incera & McLennan, 2018; Sulpizio, Del Maschio, Del Mauro, Fedeli, & Abutalebi, 2020). Moreover, there are general statistical reasons for why categorical variables are less useful analytical tools than gradient ones (Cohen, 1983; Young, 2016). While continuous measures of specific factors generally better represent language experience, categorical and/or holistic measures may nevertheless be useful for certain research goals. Here, we present various alternative approaches to operationalizing language experience that are grounded in

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

direct and gradient measures, and give examples of how to implement these analyses to predict linguistic behavior.

The current study

Our sample comprises seventy-three Hindi-Urdu speakers living in South Asia and North America, recruited online via the authors' personal and professional networks. The goal was to recruit a large sample of participants who knew both Hindi-Urdu and English; as such, recruitment materials were in English, and participants were told they would qualify for the experiment if they were over the age of 18 and able to understand Hindi-Urdu. Each participant completed an acceptability judgment task in Hindi-Urdu and filled out a language background questionnaire[1]. Note that while all participants were multilingual, we examined only Hindi-Urdu language experiences and behavior.

We first examine the extent to which (i) identifying as a Hindi-Urdu or English 'native speaker' and (ii) history of living in South Asia or North America align with more direct measures of language use and comfort. Then, we compare how well various operationalizations of language experience predict acceptability judgment behavior in two empirical domains which have the potential to be differentially affected by language experience: ratings of grammatical versus ungrammatical sentences and canonical versus non-canonical word orders. These analyses demonstrate that, compared to 'nativeness' and geography-based proxies, more direct characterizations of language experience are both more appropriate for representing language experience and more effective in linking experience to behavior.

Exploring language experience *via* clustering

This section considers how categorization of participants based on 'nativeness' and place of residence may create overly heterogeneous groups, despite being frequently-used proxies for certain (presumably homogeneous) profiles of language experience. Using exploratory clustering, we assess the degree to which NATIVE IDENTIFICATION and RESIDENTIAL HISTORY categories align with emergent language experience groups in a sample of Hindi-Urdu speakers.

Methods

Data The data for clustering comes from the language background questionnaire. Participants with incomplete responses were excluded, as was one outlier (identified during initial clustering specification) who provided a divergent pattern of ratings. Final analyses included 65 participants.

Categorical Variables Each participant was assigned to a NATIVE IDENTIFICATION (NI) and RESIDENTIAL HISTORY (RH) group. NI groups were derived from answers to two yes-or-no questions: (1) "Do you consider yourself a native speaker of Hindi-Urdu?" and (2) "Do you consider yourself a native speaker of English?". Participants were placed into one of four self-identified groups: native speaker of Hindi-Urdu (HU; n=31), native speaker of English (Eng; n=12), native speaker of both (n=15), and native speaker of neither[2] (n=8).

Similarly, RH groups were created on the basis of a pair of questions: (1) "Where did you spend the majority of your childhood?" and (2) "Where do you live now?". Locations in the U.S. or Canada were coded as "North America" while locations in India and Pakistan were coded as "South Asia". Combining answers to the two questions resulted in three main RH groups: South Asians (SouthAs; n=28), North Americans (NorthAm; n=17), and South Asia to North America migrants (SAtoNA; n=16). The small number of participants (n=5) who reported living in other locations (e.g. UAE, UK) were labeled as "uncategorized"[3].

Continuous Variables Language experience data consisted of answers to ten questions about degree of (a) childhood usage of Hindi-Urdu (hearing, speaking), (b) current usage of Hindi-Urdu (hearing, speaking, reading, writing), and (c) comfort with Hindi-Urdu (hearing, speaking, reading, writing). For each question, participants provided percentages or ratings on a sliding scale; values ranged from 0 to 100, where higher values represent more experience with Hindi-Urdu. To account for collinearity prior to clustering, highly correlated ($r \geq 0.9$) variables were averaged, resulting in a combined "Comfort with Reading and Writing" variable. All others remained independent, resulting in nine clustering variables.

Statistical Analysis To identify emergent groups with similar language experience characteristics from a bottom-up perspective, we conducted an agglomerative hierarchical clustering analysis in R (R Core Team, 2020), following recommended clustering protocols (Hair, Black, Babin, & Anderson, 2019). First, the nine language experience variables were z-scored and converted to Euclidean distances. Participants' scores were then clustered using the average linkage method, to find clusters of any size or shape. To identify a range of potential cluster solutions, the `NbClust` (Charrad, Ghazzali, Boiteau, & Niknafs, 2014) package was used to assess 30 indices which indicated 3 clusters as the majority result. Based on this, clusters of the 2-, 3- and 4-cluster solutions were profiled by inspecting distinctiveness and means; the 4-cluster solution was selected as most informative[4] and reported here.

Results

Figure 1 shows the cluster dendrogram, plotted with colored bars indicating each individual's NI and RH category. To in-

[1] Full questionnaire at `https://tinyurl.com/HUCogci2022`.

[2] A review of their questionnaires showed no other relevant language; these participants were in effect saying they did not consider themselves a native speaker of any language (cf. Rosa, 2016).

[3] Though all "uncategorized" participants currently live in a different country than where they grew up (i.e. were "migrants" like the SAtoNA group), each profile was unique; as such, we do not treat these individuals as a group. However, we retained these participants in the exploratory analysis to see how they would pattern.

[4] According to Hair et al. (2019), researcher judgment is required to determine the cluster solution that "best meets the research objectives" (p. 243), and a larger number of clusters allows for "more varied profiles", beneficial when exploring emergent patterns (p. 245).

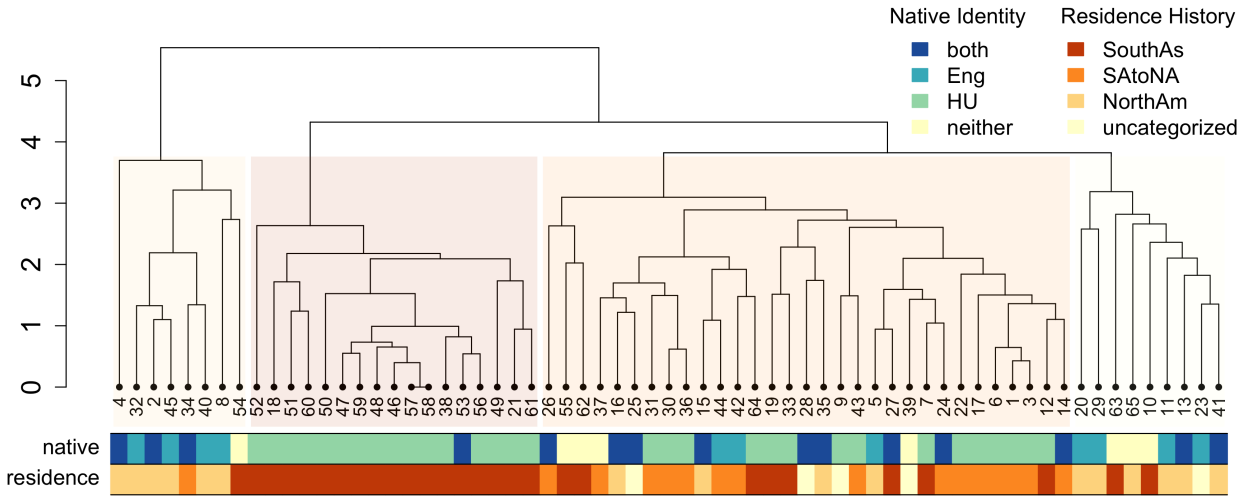

Figure 1: Dendrogram of the 4-cluster agglomerative hierarchical clustering solution, each cluster highlighted with a box (clusters A-D from left to right). Each node represents an individual, while the colored bars represent NATIVE IDENTIFICATION (top) and RESIDENTIAL HISTORY (bottom) category membership.

Table 1: Mean rating scores (0-100) for each language experience variable per cluster.

		Childhood Use		Current Use				Comfort		
Cluster	n	hear	speak	hear	speak	read	write	understand	speak	read-write
A	8	30.62	19.50	21.25	10.88	3.12	9.00	75.00	50.75	26.38
B	17	92.59	94.94	89.24	87.53	76.41	71.94	96.06	97.71	95.18
C	31	81.19	72.55	52.16	49.06	15.84	4.77	97.00	92.32	75.50
D	9	78.44	31.78	67.44	27.56	11.33	7.78	86.11	54.89	27.00

terpret the clusters, Table 1 and Figure 2 display mean ratings.

To start with the most homogeneous, all 17 individuals in Cluster B reported South Asian-only residence and identified as a native speaker of Hindi-Urdu (16 HU, 1 both). This cluster consistently provided high ratings for all aspects of language experience. In contrast, Cluster A contains a majority of North American-only individuals (6 NorthAm, 1 SAtoNA, 1 SouthAs) and most identified as a native speaker of English (4 Eng, 3 both, 1 neither); notably, none identified as only a Hindi-Urdu native speaker. This cluster has scores generally in the low range, as well as the lowest relative comfort. These two clusters appear to represent the higher and lower ends of Hindi-Urdu language experience, which broadly align with prototypical NI and RH expectations.

On the other hand, Clusters C and D have middling scores and present mixed bags in terms of RH and NI. While Cluster C contains nearly all of the migrants (15 SAtoNA, 4 uncategorized), it also includes participants of the other two categories (8 SouthAs, 3 NorthAm). Native identification is similarly mixed, with all four labels represented. This cluster is characterized by moderately high childhood use paired with lower current use but high comfort. Finally, Cluster D involves a combination of North American- and South Asian-only residents (6 NorthAm, 2 SouthAs, 1 uncategorized) with non-Hindi-Urdu only native identification (4 Eng, 3 neither, 2 both). This cluster shows a distinctive pattern of moderately high listening experience with lower speaking and reading/writing experience.

Overall, while we do find some consistent patterning between NI, RH, and language experience measures (namely in Clusters A and B), these different variables do not fully or neatly align. The corrected Rand Index (in which values closer to 1 represent a better match of clusters to external categories) confirms that RH is better matched to these clusters (Rand=0.249) than NI (Rand=0.178).

Summary

Hierarchical clustering identified four emergent clusters of Hindi-Urdu speakers with distinct language experience profiles based on CHILDHOOD USE, CURRENT USE, and COMFORT. RH provides a better approximation than NI for these Hindi-Urdu language profiles, but, as expected, neither category type explains cluster composition on the whole.

For instance, although a subset of South Asians who identify as Hindi-Urdu native speakers *did* form a homogeneous language experience cluster, roughly one-third of South

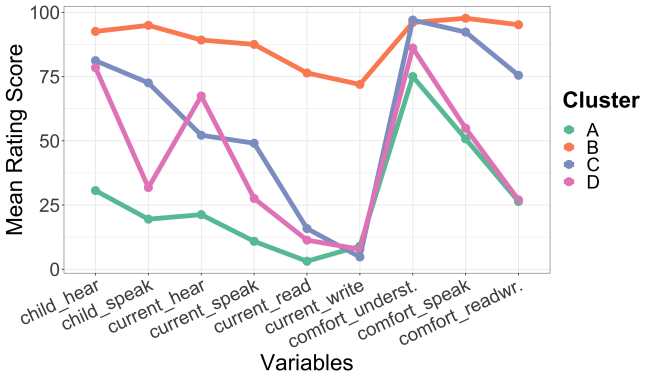

Figure 2: Mean rating scores (0-100) for each language experience variable. Each line represents one cluster.

Asians and half of Hindi-Urdu native-identifying speakers did *not* show similar patterns of language experience ratings. Based on this sample, this result suggests that if either categorical proxy were used as a grouping variable, a substantial number of individuals' language experiences would not be well-represented—either by the group average or by the researchers' assumptions of their language experience.

Exploring the predictivity of different experience measures *via* Bayesian analysis

Although we did not find a direct connection between NI or RH categories and the gradient language experience measures we collected, it could be that these proxies are useful for explaining variability in linguistic behavior. For example, these proxies may tap into important aspects of language experience that we did not measure directly, but which do in fact affect language use, such as the sense of allegiance to a language that might come from identifying as a 'native speaker'. This section asks how different operationalizations of language experience perform in predicting one kind of language behavior: acceptability judgments.

We first investigate how these measures map onto ratings of grammatical and ungrammatical sentences. Previous work has shown that *ex situ* speakers (as well as language learners) are more likely to rate ungrammatical sentences as being relatively acceptable; that is, they are less likely to reject sentences that would be rejected by *in situ* participants. This has been explained via various mechanisms; notably, the metalinguistic nature of the task has caused researchers to connect this pattern to discomfort or lower confidence (González-Vilbazo et al., 2013; Orfitelli & Polinsky, 2017). As such, an ideological measure such as self-identification as 'native' could potentially predict differences in acceptability, insofar as it could be a proxy for comfort. So, in addition to asking about the role of RH and NI, we also ask how well COMFORT, measured gradiently, might predict acceptability ratings.

We then investigate how well language experience measures and their proxies predict ratings of grammatical sentences which differ in how frequent and tied to discourse they are. We compare acceptability ratings of Hindi-Urdu sentences in the canonical Subject-Object-Verb (SOV) order to sentences in Subject-Verb-Object (SVO) order.[5] While SVO order is grammatical in Hindi-Urdu, it is neither discourse-neutral nor as frequent as SOV (Manetta, 2012; Patil et al., 2008); however, SVO order is predominant in English, which is the other relevant language for these participants. Previous work found that language experience shapes ratings of such sentences (Namboodiripad, Kim, & Kim, 2019; Anderssen, Lundquist, & Westergaard, 2018), but these studies used categorical proxies such as residential history to divide participants. Here, we ask whether gradient measures would indeed better capture language experience-based variation.

Methods

We use a Bayesian approach to determine which predictors provide the best fit to the data. There are two main reasons to prefer Bayesian models to traditional linear regressions in this case. First, in order to (easily) compare linear regressions, they must use the same parameters or a subset thereof; here, we need to compare models with different sets of predictors. Second, continuous predictors will generally account for more variance than categorical predictors. We must thus determine whether any improvement is more than we would expect simply from replacing a categorical variable with a continuous one. Bayesian models that use different sets and types of predictors can be directly compared using leave-one-out (LOO) cross-validation, which provides a measure of fit that corrects for potential issues such as over-fitting (Vehtari, Gelman, & Gabry, 2017).

Data The same 65 participants from the clustering analysis were included in this analysis. Participants heard and rated 93 sentences on a 1–7 Likert scale. Items were recorded by a speaker of Hindi-Urdu and are available at `https://tinyurl.com/HUCogci2022`. Participants were asked to rate each sentence based on how it sounded to them as a user of Hindi-Urdu. As motivated above, two subsets of the stimuli were analyzed: (1) 19 ungrammatical and 20 grammatical sentences and (2) 5 SOV and 5 SVO sentences.

Predictor Variables For each subset of rating data, we compared five main statistical models that varied in the nature and categoricity of their predictors: Two based on the *a priori* categorical proxies (NI and RH), one based on the clusters inferred from the previous analysis, and two based on composite continuous variables—a simple additive score on the one hand, and component scores inferred through Principal Component Analysis (PCA) on the other hand.

The additive score was calculated as an unweighted linear combination of the ten continuous language experience variables from the language background questionnaire. To obtain component scores, a PCA was conducted on these same variables. Parallel analysis indicated that the optimal num-

[5]Items consisted of animate subjects, inanimate objects, and transitive verbs.

Table 2: LOO results: Grammaticality

Model	*elpd_diff*	*SE_diff*
Comfort	0.0	0.0
Clustering	−0.2	7.5
PCA	−4.9	6.1
Additive score	−17.0	7.3
Childhood use	−20.2	9.5
Native identification	−29.5	10.8
Residential history	−36.7	10.3
Current use	−46.0	11.3

Table 3: LOO results: Word order

Model	*elpd_diff*	*SE_diff*
Additive score	0.0	0.0
Current use	−0.5	1.5
PCA	−0.6	1.5
Childhood use	−1.1	1.9
Clustering	−1.6	2.3
Native identification	−2.5	4.3
Residential history	−3.2	3.6
Comfort	−3.4	2.1

ber of factors was two. However, for consistency with the other models, only the first principal component (PC1) was included in the PCA-based model. This dimension involved high loadings on all variables except current reading and writing, with particular weight placed on childhood use and current comfort. Though it is not transparent, this weighted composite variable can be seen as a measure capturing these various aspects of language experience holistically.

We fit an additional three models exploring the degree to which the specific predictors COMFORT, CHILDHOOD USE, and CURRENT USE of the language contributed to model fit. In contrast to the "holistic" nature of the predictors used in the first set of models, these predictors can be thought of as "specific" (i.e. targeting specific aspects of language experience rather than experience as a whole).

Statistical Analysis Models were constructed with the following general form, where MEASURE stands in for each of the eight predictors we tested, and CONDITION stands in for either the grammaticality (grammatical vs. ungrammatical) or word order (SOV vs. SVO) conditions:

$$\text{rating} \sim \text{MEASURE} * \text{CONDITION} + (1|\text{PARTICIPANT}) + (1|\text{ITEM})$$

Models were fit using `Stan` via the `brms` package in R (Bürkner, 2021). We chose weakly informative priors, using $normal(0,1)$ for intercepts and slopes, and $cauchy(0,1)$ for standard deviations. For each model, two independent MCMC chains were run for 5,000 iterations, with a warm-up phase of 1000 iterations.

Results

Interpretation The difference in ELPD (*elpd_diff*) is the difference in the expected log pointwise predictive densities (ELPD) of a pair of models. The standard error (SE) represents the uncertainty around *elpd_diff*. In interpretation, it is customary to consider how much larger *elpd_diff* is than the corresponding SE, as these SEs do not provide confidence intervals around *elpd_diff*. Here, we treat two times the SE as a lower bound such that if an *elpd_diff* value is greater than double the corresponding SE, we consider that model to be reliably different from the baseline model (i.e., the model with zero *elpd_diff* and SE). In addition, SE is poorly estimated for small datasets ($N < 100$). We thus interpret the ordering of the models with the caveat that replication, ideally on a larger dataset, should be used to confirm these patterns.

Grammaticality The results of model comparison via LOO cross-validation are given in Table 2. The model based on COMFORT alone was the most predictive of judgments of (un)grammatical sentences. In addition, the models using CLUSTERING and PCA—both of which take into account comfort but also reflect other aspects of language use—did not strongly differ from the COMFORT model, indicating that they were similarly predictive. Conversely, the proxy models (NI and RH) were near the bottom, along with CURRENT USE, which provided the least-good fit to the data relative to baseline. Overall, the holistic continuous measures and the inferential measures outperformed *a priori* categorical ones. Notably, while COMFORT was the best predictor of how participants rated these sentences, neither NI or RH, which are reasonable candidates for being indirect proxies of linguistic comfort, performed similarly.

Word Order The results of LOO cross-validation are given in Table 3. Here, no model resulted in an *elpd_diff* greater than 4, which is the standard lower bound for a meaningful difference in ELPD between models. This is likely due to the fact that all orders were grammatical; as a result, there was a smaller difference in ratings between orders. We therefore treat these results as inconclusive, though we provide a description and tentative interpretation of their ordering, pending replication with more data. The ADDITIVE SCORE and CURRENT USE models appeared to be the top performers, while COMFORT and the NI and RH proxies provided a relatively poorer fit. That the ADDITIVE SCORE and CURRENT USE models potentially perform best suggests that language experience as a whole, potentially driven by language use, may be most relevant to word order judgments (as opposed to, for example, language comfort).

Summary

LOO cross-validation of a set of Bayesian models found that the *a priori*, categorical, and holistic variables NATIVE IDENTIFICATION and RESIDENCE HISTORY were not

strong predictors of acceptability judgment behavior. They were outperformed by (a) inferential holistic variables (PCA, CLUSTERING), (b) continuous holistic variables (ADDITIVE SCORE) and (c) certain *a priori* but continuous and specific variables (COMFORT, CURRENT USE). These proxies do not approximate language experience well, nor do they obviously contribute additional insight; use of almost any other factor would result in finding a stronger relationship between language experience and behavior in this data.

This result highlights the potential issues with in using native identification or residential history as selection criteria for participants. In addition, we see that different *kinds* of language experience best predict different kinds of linguistic behavior. In this sample, we found that comfort best predicts judgments of grammatical and ungrammatical sentences, while an additive measure of lifetime language experience best predicts judgments of different (grammatical) word orders in Hindi-Urdu. Thus, we demonstrate the importance of identifying and testing specific experience factors that might contribute to linguistic behavior rather than relying on *a priori*, holistic, and categorical proxies.

Discussion & Conclusion

This paper presented two analyses demonstrating that, in a population of Hindi-Urdu speakers, common proxies for language experience (NATIVE IDENTIFICATION, RESIDENTIAL HISTORY) neither *accurately* nor *adequately* represented our chosen facets of language experience (CHILDHOOD USE, CURRENT USE, COMFORT). First, they did not accurately map onto contextually-relevant language experience profiles identified via clustering analysis. Second, they did not adequately approximate the relationship between language experience and behavior (here, acceptability judgments), evidenced by their poor predictive performance relative to other measures in a model comparison.

Certainly, this study only investigated a few facets of language experience; many others could have plausibly been included (e.g. age of acquisition). Further, the particular degree of mapping between proxies and more direct measures are expected to vary depending on the research population and context. The point remains: Researchers cannot assume that these proxies will align well with language experience without confirming directly for their particular sample. Despite their continued prevalence in psycholinguistic research, our results thus explicitly challenge the utility of 'nativeness' and residence as proxies for language experience, both conceptually and methodologically. We also show that effectively operationalizing language experience requires direct and gradient assessments of particular facets of language experience (e.g. Luk & Bialystok, 2013), even when only considering one language. For researchers considering this approach, we suggest three ways one could handle such multi-dimensional data, depending on the research questions and goals.

Some researchers may be using categorical groups because of a specific interest in profiles of language experience. We show that top-down groups based on assumed profiles do not map to emergent profiles, at least in our sample. If particular combinations of experience factors are relevant, we recommend clustering on gradient measures of language experience to find the profiles actually represented in the data, then use those in analyses of language behavior; the clusters which emerged in our study did well compared to the alternatives at predicting how (un)grammatical sentences were rated.

However, we expect that many cognitive scientists are instead interested in the mechanisms underlying language behavior. For questions which are general or exploratory in nature (e.g. does overall language experience play a role?), researchers can use multiple gradient measures assessing various aspects of experience to create a variable that encompasses language experience holistically. One option is using a data-driven approach, such as PCA, to find the most informative dimensions to represent the variability in the data. In this case, PC1 represented many factors, notably both usage and comfort, and was able to reveal a role of language experience in predicting ratings of (un)grammatical sentences (and possibly SOV versus SVO sentences). Another way we did this was an additive score (cf. Gertken, Amengual, & Birdsong, 2014), which in our data may represent sheer exposure/usage; the interpretation of such a variable will be dependent on the measures collected by the researcher. Nevertheless, we show that this holistic gradient measure performed relatively well in predicting acceptability judgments for grammaticality (as well as potentially very well for word orders, which may be subject to language contact effects). Using either type of composite variable is a viable strategy.

Alternatively, if there is reason to narrow down the research question to particular facets of language experience, one can default to a researcher-driven approach to identify conceptually-relevant variables. For example, comfort was hypothesized to be particularly relevant for rejecting ungrammatical sentences (González-Vilbazo et al., 2013). Our combined COMFORT score ended up being the most predictive of how participants judged the acceptability of grammatical versus ungrammatical sentences, confirming our hypothesis.

To conclude, we do not argue that language researchers should never use variables like 'nativeness' and residential history. Rather, we propose that researchers carefully consider (i) which aspects of language experience are relevant factors in their studies and (ii) how neatly these factors may or may not map onto the simplifications that they are using. For example, questions about 'nativeness' are in fact always questions about *ideologies of nativeness*; thus, if the factors of interest are not the ideologies themselves, the use of other variables would be more appropriate. Given the complexities, the simpler option—conceptually and methodologically—will often be to directly measure what we want to measure. By arguing for such an approach and providing a demonstration of how to implement it, we hope to contribute to a shift away from vague proxies and towards more explicit models of how language experience shapes cognitive processes.

Acknowledgments

LSPC is supported in part by funding from the Social Sciences and Humanities Research Council of Canada. We are thankful to the anonymous reviewers for their feedback and to Jonathan Brennan for advice on our analyses.

References

Anderssen, M., Lundquist, B., & Westergaard, M. (2018). Cross-linguistic similarities and differences in bilingual acquisition and attrition: Possessives and double definiteness in norwegian heritage language. *Bilingualism: Language and Cognition*, *21*(4), 748–764.

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. doi: 10.18637/jss.v100.i05

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36.

Cheng, L. S. P., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology*, *12*.

Cohen, J. (1983). The Cost of Dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. doi: 10.1177/014662168300700301

Debenport, E. (2011). As the rez turns: Anomalies within and beyond the boundaries of a pueblo community. *American Indian Culture and Research Journal*, *35*(2), 87–110.

Dewaele, J.-M. (2018). Why the Dichotomy 'L1 Versus LX User' is Better than 'Native Versus Non-native Speaker'. *Applied Linguistics*, *39*(2), 236–240. doi: 10.1093/applin/amw055

Dewaele, J.-M., Bak, T. H., & Ortega, L. (2021). Why the mythical 'native speaker' has mud on its face. In N. Slavkov, N. Kerschhofer-Puhalo, & S. M. Melo-Pfeifer (Eds.), *Changing face of the "native speaker": Perspectives from multilingualism and globalization.* Mouton De Gruyter.

Faez, F. (2011). Reconceptualizing the native/nonnative speaker dichotomy. *Journal of Language, Identity & Education*, *10*(4), 231–249.

Fernández-Dobao, A., & Herschensohn, J. (2021). Acquisition of spanish verbal morphology by child bilinguals: Overregularization by heritage speakers and second language learners. *Bilingualism: Language and Cognition*, *24*(1), 56–68.

Gertken, L. M., Amengual, M., & Birdsong, D. (2014). 11. assessing language dominance with the bilingual language profile. In *Measuring l2 proficiency* (pp. 208–225). Multilingual Matters.

González-Vilbazo, K., Bartlett, L., Downey, S., Ebert, S., Heil, J., Hoot, B., … Ramos, S. (2013). Methodological considerations in code-switching research. *Studies in Hispanic and Lusophone Linguistics*, *6*(1), 119–138.

Gullifer, J. W., Kousaie, S., Gilbert, A. C., Grant, A., Giroud, N., Coulter, K., … Titone, D. (2021). Bilingual language experience as a multidimensional spectrum: Associations with objective and subjective language proficiency. *Applied Psycholinguistics*, *42*(2), 245–278.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (Eighth edition ed.). Andover, Hampshire: Cengage.

Incera, S., & McLennan, C. T. (2018). Bilingualism and age are continuous variables that influence executive function. *Aging, Neuropsychology, and Cognition*, *25*(3), 443–463.

Kim, J. Y. (2020). Discrepancy between heritage speakers' use of suprasegmental cues in the perception and production of spanish lexical stress. *Bilingualism: Language and Cognition*, *23*(2), 233–250.

Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology*, *25*(5), 605–621.

Manetta, E. (2012). Reconsidering rightward scrambling: Postverbal constituents in hindi-urdu. *Linguistic Inquiry*, *43*(1), 43–74.

Marian, V., & Hayakawa, S. (2021). Measuring bilingualism: The quest for a "bilingualism quotient". *Applied Psycholinguistics*, *42*(2), 527–548.

Namboodiripad, S., Kim, D., & Kim, G. (2019). English dominant korean speakers show reduced flexibility in constituent order. *Proceedings of CLS*, *53*.

Orfitelli, R., & Polinsky, M. (2017). When performance masquerades as comprehension: Grammaticality judgments in experiments with non-native speakers. In *Quantitative approaches to the russian language* (pp. 197–214). Routledge.

Ortega, L. (2020). The study of heritage language development from a bilingualism and social justice perspective. *Language Learning*, *70*, 15–53.

Patil, U., Kentner, G., Gollrad, A., Kügler, F., Féry, C., & Vasishth, S. (2008). Focus, word order and intonation in hindi. *Journal of South Asian Linguistics*, *1*.

R Core Team. (2020). *R: A language and environment for statistical computing.* Vienna, Austria.

Rampton, M. B. H. (1990). Displacing the 'native speaker': Expertise, affiliation, and inheritance. *ELT Journal*, *44*(2), 97–101. doi: 10.1093/eltj/44.2.97

Rosa, J. D. (2016). Standardization, racialization, languagelessness: Raciolinguistic ideologies across communicative contexts. *Journal of Linguistic Anthropology*, *26*(2), 162–183.

Sulpizio, S., Del Maschio, N., Del Mauro, G., Fedeli, D., & Abutalebi, J. (2020). Bilingualism as a gradient measure modulates functional connectivity of language and control networks. *NeuroImage*, *205*, 116306.

Tsehaye, W., Pashkova, T., Tracy, R., & Allen, S. E. M.

(2021). Deconstructing the native speaker: Further evidence from heritage speakers for why this horse should be dead! *Frontiers in Psychology*, *12*. doi: 10.3389/fpsyg.2021.717352

Unsworth, S. (2019). Quantifying Language Experience in Heritage Language Development. In *The Oxford Handbook of Language Attrition* (pp. 433–445). Oxford University Press. doi: 10.1093/oxfordhb/9780198793595.013.34

Uygun, S., Schwarz, L., & Clahsen, H. (2021). Morphological generalization in heritage speakers: The turkish aorist. *Second Language Research*, 02676583211059291.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

y Cabo, D. P. (2020). Examining the role of cross-generational attrition in the development of spanish as a heritage language: Evidence from gustar-like verbs. *Linguistic Approaches to Bilingualism*, *10*(1), 86–108.

Young, M. E. (2016). The problem with categorical thinking by psychologists. *Behavioural Processes*, *123*, 43–53. doi: 10.1016/j.beproc.2015.09.009