Studies in Second Language Acquisition (2023), **45**, 1186–1209
doi:10.1017/S0272263123000177

RESEARCH ARTICLE

Variability and individual differences in L2 sociolinguistic evaluations: The GROUP, the INDIVIDUAL and the HOMOGENEOUS ENSEMBLE

Mason A. Wirtz[1] and Simone E. Pfenninger[2]

[1]University of Salzburg, Department of German Language and Literatures, Salzburg, Austria; [2]University of Zurich, English Department, Zurich, Switzerland
Corresponding author: Simone E. Pfenninger; Email: simone.pfenninger@es.uzh.ch

(Received 06 August 2022; Revised 28 February 2023; Accepted 03 March 2023)

Abstract

This study is the first to investigate subject-level variability in sociolinguistic evaluative judgements by 30 adult L2 German learners and explore whether the observed variability is characterizable as a function of individual differences in proficiency, exposure, and motivation. Because group-level estimates did not paint an accurate picture of the individual, we propose methods capable of integrating population-level estimates with person- and ensemble-centered approaches so as to reconcile generalizability and individuality. Using random effects from Bayesian mixed-effects models, we found that global subject-level variability in evaluative judgements was not predicted by individual differences. By building homogeneous ensembles (i.e., subgroups of individuals with similar evaluative judgements), however, it was possible to assess whether ensembles were characteristic of certain levels of individual differences. This ensemble-centered approach presents an innovative way to address the *group-to-individual generalizability* issue in cross-sectional data and transcend individual variability in order to make tentative generalizations of individual cases to wider populations.

Introduction

Research in variationist SLA has repeatedly positioned the acquisition of sociolinguistic competence as an individually owned process (Ender, 2019; Howard, 2012; Kinginger 2008; Regan, 2010; van Compernolle & Williams, 2012) in both production and perception. Given this axiom, it is problematic to base conclusions exclusively on group estimates, as these tend to obscure the learning paths and patterns characteristic of the *individual*. This places a premium on designs and statistical models that are able to closely monitor how the individual behaves in relation to the group and vice versa—that is, studies that include variability measures that are able to provide additional information to

that of a mean performance.[1] In so doing, the field of variationist SLA (and beyond) can continue to innovatively tackle the challenge of developing and refining methods for integrating quantitative, population-level estimates with person- and/or ensemble-centered approaches so as to reconcile generalizability and predictive power with individuality and variability. One possible way to approach this is to make better use of the random effects in mixed-effects models, and we demonstrate in this contribution one potential method to strategically employ random effects as operational measures of individuality—that is, subject-level variability. We focus specifically on random effects because these (a) have become common methods in nearly every linguist's toolkit and (b) can help measure the relation between group, subgroup, and individual.

The present contribution takes a complexity theory perspective (Larsen-Freeman & Cameron, 2008) to analyze second language (L2) learners' sociolinguistic evaluative judgements in Bavarian-speaking Austria, with a focus on intelligence and friendliness judgements of standard German and dialect,[2] respectively. The rationale behind our focus on friendliness and intelligence judgements stems from previous research asserting that these are common socioindexical attributions for the standard German and dialect varieties in Bavarian-speaking Austria[3] (e.g., Bellamy, 2012; Soukup, 2009). Given the extensive subject-level variability (i.e., the extent to which subjects deviate from the group-level estimate) in the data, group estimates did little in the way of painting an accurate picture of the individuals within the group. Because of this, we propose an exploratory approach to address highly heterogeneous data sets. Although the data set, as will be elucidated in the methods section, is based on a limited number of participants (30 subjects) and verbal stimuli (four judgements per each participant), our goals are strongly methodologically oriented. In our analyses, we first address whether subject-level variability can be systematically predicted by (a constellation of) individual differences in varietal proficiency, exposure, and motivation and then identify whether certain levels of the aforementioned individual differences are more characteristic of particular subgroups of similarly behaving individuals (i.e., homogeneous ensembles) as a way to transcend the individual heterogeneity in the data. In delivering this explication, we have three interrelated goals: first, to home in on the subject-level variability present in sociolinguistic evaluative judgement data as well as any predictors thereof; second, to propose methods for quantifying and assessing subject-level variability that can be easily integrated into variationist SLA scholars' toolkits; and, third, to tentatively apply methods that can—even with cross-sectional data—reconcile and operationalize generalizability with heterogeneous groups, an issue that will have implications for future variationist SLA research and foreign/second language practitioners so as to better inform and do justice to the individually owned process of sociolinguistic development.

Acquisition of sociolinguistic competence and the obsession with the group

The field of variationist sociolinguistics has long employed quantitative methods to assess linguistic and extralinguistic constraints probabilistically guiding patterns of

[1]Of course, we make note that variability and the issue of the group vs. individual is by no means exclusive to SLA and has been grappled with for decades in related fields such as variationist sociolinguistics (e.g., Guy, 1980; Tagliamonte, 2006).

[2]Note that in German-speaking sociolinguistics, the term "dialect" is used in the spirit of "local base dialect" or "local vernacular" rather than synonymous to "any language variety."

[3]Note that we employ the term *Bavarian* in its dialectological sense. It refers to eastern varieties of Upper German, which are spoken in much of Austria (thus, Bavarian-speaking *Austria*).

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

variation commonly found in human language. These variationist techniques have since been adopted to examine how L2 learners acquire sociolinguistic competence—that is, "the capacity to recognize and produce socially appropriate speech in context" (Lyster, 1994, p. 263; for an overview, see Geeslin & Long, 2014; Howard et al., 2013; Regan et al., 2009). Although the majority of these studies has investigated learner intra-/interspeaker variation, there is amassing research on the perception end of sociolinguistic competence, focusing specifically on learners' attitudes toward and attachment of meaning to specific variants/varieties (for German: Ender, 2020; Ender et al., 2017; Kaiser et al., 2019; for English: Clark & Schleef, 2010; Davydova et al., 2017; McKenzie, 2008; for Spanish: Chappell & Kanwit, 2022; Geeslin & Schmidt, 2018; Schmidt & Geeslin, 2022). As Kanwit and Geeslin (2020) underscore, the ability to interpret additional meaning communicated via language variation is essential for successful communication and thus a central dimension of sociolinguistic competence. This holds true for the Bavarian-Austrian landscape (Kaiser et al., 2019) such that L2 learners are challenged to acquire the socioindexical attributions associated with language variation commonly held by the target language community. This includes the notion that "there are certain things that one can and cannot do with *either one* of the varieties" (Soukup, 2009, p. 128, italics in original). For example, informants in Bavarian-speaking Austria have been shown to attribute dialect strong affective value with respect to naturalness, honesty, emotionality, likeability, relaxedness, and humor. However, dialects cannot typically project characteristics related to intelligence, education, politeness, seriousness, and refinement in the way the standard German variety can (Bellamy, 2012; Soukup, 2009). Such systematic and context-bound varietal usage obliges learners to acquire knowledge on dialect's association with covert social prestige (friendliness, localness, etc.) and standard German's common projection of intelligence and education. That is, L2 learners' ability to understand, interpret, and decode such subtle indicators of social and situational information can be a deciding factor in adeptly "reading between the sociolinguistic lines" and thus successfully participating in social, commercial, and academic interactions.

Ender et al. (2017) and Kaiser et al. (2019) addressed whether adolescent and adult L2 learners ascribe standard and dialect varieties in Austria similar ideological and social meaning as does the L1 community. Early results suggest that L2 learners attribute dialect more critical evaluations compared with the standard German stimuli. L2 learners also rated the dialect stimuli as a whole more negatively than did the native speakers in the study. In contrast, the native and L2 speakers' attitudinal patterns regarding the standard German stimuli appeared similar, with slight tendencies indicating that the L2 learners attributed higher value to the standard variety than did the native speakers. In a more systematic interindividual analysis, Ender (2020) found in the combined sample of the aforementioned native and L2 speakers that dialect proficiency appeared predictive of differences in evaluative judgements, whereas the L1 versus L2 binary did not meaningfully explain the variance in evaluations.

By and large, both production- and perception-based studies have employed traditional quantitative variationist methods (i.e., some form of regression model) so as to ascertain the effect of the predictor variables in question. Within these quantitative approaches, mixed-effects models have been gaining steady momentum (see, e.g., Gudmestad et al., 2020), specifically to account for nonindependence and autocorrelation in the data arising from subject-level variability. This subject-level variability is oftentimes included as a random effect in such models but subsequently neglected in the interpretation of the data in lieu of focusing on the population-level effects. This strict focus on population-level effects, however, has the unequivocal potential to lead

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

astray, given the axiom that L2 learners vary idiosyncratically in their second language use and (can) deviate rather drastically from group norms of variation. This makes it essential to additionally scrutinize individual norms of variation "so as not to obscure individual differences through the reporting of only group norms of use" (Geeslin et al., 2013, p. 156; see also, Howard, 2012).

Group-level effects versus individual variability

In variationist SLA, the role of (particularly linguistic and social) individual differences has been a central point of inquiry. With a particular eye on the perception side of sociolinguistic competence, more target-like L2 sociolinguistic evaluative judgements have been associated with (a) quantity and quality of contact with the L2 variation landscape and/or input from and experiences with target language speakers in the L2 environment (e.g., Chappell & Kanwit, 2022; Davydova et al., 2017; Geeslin & Schmidt, 2018), (b) language proficiency more generally (e.g., Chappell & Kanwit, 2022; Davydova et al., 2017), and (c) varietal proficiency—for example, proficiency in Austro-Bavarian dialect versus standard German (Ender, 2020). However, research on the acquisition of sociolinguistic variation remains ripe for results on two fronts: First, it is necessary to better disentangle how learners' relations to *individual varieties* such as dialect and standard German differently affect their perception of said varieties. This can more generally aid in facilitating a clearer understanding concerning the extent to which learners integrate different varieties into their multivarietal repertoires based on their explicit experiences with, exposure to, and proficiency in the respective varieties. Second, as Geeslin et al. (2013) and Kanwit (2022) note, variationist SLA research must more aggressively address how socioaffective factors such as motivation affect L2 learners' sociolinguistic repertoires. To our knowledge, the results in George (2014) provide the lone exploration concerning the extent to which quantitatively captured motivational factors explain differential outcomes in sociolinguistic competence, though her results focused on differences in production in L2 Spanish. Notably, motivation did not explain higher frequencies of uses of the sociolinguistic variants under scrutiny in this case. Nonetheless, it stands to reason that motivation can be associated with, for example, more target-like evaluative judgements. This is because differences in socioaffective factors may affect "initial orientation toward learning as well as changes in perspective over time" (Geeslin & Schmidt, 2018, p. 389) and also learners' evaluative orientations toward a particular language and its varieties.

Although quantitatively capturing systematic differences in particular factors and regressing language production/perception against these remains the prevailing approach for assessing the effects of individual differences, there exists a growing call to focus on the individual learner or smaller groups thereof—for example, in the spirit of qualitative, person-in-context, and/or individual-level analyses—in addition to providing group estimates (for recent discussions concerning the need to consider individual learner data and applications of this, see Kanwit, 2017, 2019). This is because person-centered analyses can provide complementary (and often more nuanced) insights into differences in developmental trajectories and/or motives for acquisition. Howard (2012), for example, called into question previously widely accepted results (e.g., classroom learners making minimal use of informal variants) by focusing the contextual lens on individual learners and their progression over time as well as which internal and external variables affect such development. Further qualitative, person-in-context-based explorations into sociolinguistic development have moreover gone to

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

show the bandwidth of individualism in learners' sociolinguistic repertoires (Ender, 2019; Kinginger, 2008; van Compernolle, 2019; van Compernolle & Williams, 2012). These qualitative investigations urge a more concentrated focus on the individual, their patterns of sociolinguistic development and how these are either confluent with or diverge from the group. In applied linguistics, the theory of complex dynamic systems, which provides a general framework for studying change on various time scales, has inspired a continuously growing wave of L2 research to steer away from the traditional practice of relying solely on the statistical means of (group) performance and focusing more on the individual and the scales of variability that come with this (Kliesch & Pfenninger, 2021; Lowie & Verspoor, 2019).

As a whole, the issue regarding the relationship between group estimates and individual behavior is not new and has been a matter of general interest in variationist sociolinguistics (with respect to L1 speech communities) for decades (see, e.g., Guy, 1980; Tagliamonte, 2006). As Regan (2004, p. 340) puts it, however, "researchers with a variationist approach have tended to group learners together," oftentimes extrapolating that the community patterns assumed to hold in variationist studies of first language use transfer when examining L2 learners' acquisition of variation in the L2. Regan (2004) explored this phenomenon of individualism in terms of learners' differential acquisition of sociolinguistic competence as compared with that of the group. In her preliminary study with $n = 5$ participants, she found similar patterns of French sociolinguistic *ne* deletion in group and individual, showing that individual patterns of variation in the acquisition of a certain variant might closely match group patterns. Bayley and Langman (2004) similarly investigated group patterns of variation as opposed to how the individual behaves, albeit with a focus on native versus nonnative variation, also underscoring that L2 learners' individual patterns of acquisition of verbal morphology in English and Hungarian (sample size $n = 20$) closely matched those of the group. Such findings from a variationist viewpoint are, of course, encouraging, as they would posit that group estimates provide an accurate picture of individual variation behavior in use. As Geeslin et al. (2013) lamented, however, the aforementioned investigations strictly considered either one or two proficiency levels, thus obscuring how individual variability shifts (i.e., grows, declines) dynamically across proficiency levels. Geeslin et al. (2013) addressed this issue in more detail and underscored that individual norms of variation are, indeed, prone to change across, for example, proficiency levels such that variability decreases with increasing proficiency, though it does not disappear entirely.

The issue of ergodicity

The dynamic turn in SLA from considering within-task variability as vulnerability, measurement error, or unnecessary noise (e.g., Bülow & Pfenninger, 2021) to the observation that intraindividual variation may be better representative of the level of the individual than means-based analyses is in part due to a revelation of the drawbacks of the so-called ergodicity hypothesis. The principles of ergodicity state that analyses of interindividual variation must yield the same result as those of intraindividual variation so as to meet the conditions of generalizability (Molenaar, 2015). In other words, "we cannot generalize group statistics—especially when we deal with human beings—to the individual, and vice versa, unless the group is an ergodic ensemble" (Lowie & Verspoor, 2019, p. 185). To be considered an ergodic ensemble, two stringent conditions must be met: (a) the homogeneity criterion (processes are equivalent for group and individual)

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

and (b) the stationarity criterion (the mean and variance of the process[es] remain stable over time). Such strict conditions, however, impose on the very research traditions regarding generalizability practiced today such that nearly every group of learners, speakers, or *human beings* as a whole is nonergodic. Strictly speaking, then, hardly any sampling procedures or designs can produce an ergodic ensemble from which generalizations to a broader population can be extrapolated.

This is problematic inasmuch as one predominant goal of applying inferential statistics is to be able to generalize a set of observations to a wider population that the same sampling procedure would produce (i.e., external validity). However, drawing such population-level conclusions is oftentimes difficult to achieve, particularly in research focusing on the acquisition of variation and in SLA as a whole, given the tremendous individual variation in L2 learning outcomes and, by extention, the inherent heterogeneity of acquisition and development (Dörnyei, 2006; Ellis, 2004; Geeslin et al., 2013). The field of applied linguistics has begun addressing such issues, particularly within the meta-theoretical framework of complex dynamic systems theory (CDST; for a recent review of the CDST research in SLA, see Hiver et al., 2022), and there are increasing calls for innovative methods to more accurately transfer findings to wider populations (see, e.g., Lowie & Verspoor, 2019; Peng, Lowie, & Jager, 2022).

For example, it has been suggested that adopting a starker person-centered, individual approach could allow us to identify subgroups of similarly behaving individuals within the data, which could in turn functionally be treated as (ergodic) ensembles (Lowie & Verspoor, 2019; Molenaar, 2015; Molenaar & Campbell, 2009; Peng, Lowie, & Jager, 2022). This approach, then, allows "the findings at the subgroup level and those of the individuals composing the subgroup [to be] mutually inferable" (Peng, Lowie, & Jager, 2022, p. 894). Of course, it must be noted that identifying true ergodic ensembles is only possible when using dense time-serial measurements, as only so can the stationarity criterion be justified and met (which may be a goal of future longitudinal designs). This, however, is not an indication that cross-sectional studies such as this one should not similarly attempt to elucidate learner heterogeneity—quite the contrary. Against the backdrop of a cross-sectional design, Peng, Jager, and Lowie (2022) similarly advocate for a bottom-up approach to identify emergent patterns arising from the data so as to identify more predictable manifestations of individual variety. This allows us to identify distinct homogeneous subgroups displaying similarities in the processes observed. We term such subgroups *homogeneous ensembles* (or ensembles, for short), given that, although these subgroups cannot unequivocally be deemed *ergodic* (stationarity criterion cannot be tested), these ensembles are a possibility to fulfill one of the stringent conditions for ergodic ensembles with cross-sectional data. In so proceeding, it should then be possible for cross-sectional data sets to transcend the individual variety and to make "careful generalization[s] of individual cases" (Peng, Jager, & Lowie, 2022, p. 3).

The present study

The current study is part of a broader project that seeks to explore the role of linguistic, socioaffective, and cognitive factors in sociolinguistic development in adult L2 German learners, with respect to both production and perception. This study is novel in several respects: To our knowledge, it is the first investigation that systematically explores (a) the extent to which L2 learners' patterns of sociolinguistic judgements deviate from those of the group, (b) whether subject-level variability can be explained by individual differences, and (c) whether subgroups of similarly behaving individuals are similarly

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

influenced by individual differences factors. The results should shed light on the extent of individual variability in L2 acquisition of sociolinguistic variation (i.e., how does the individual compare with the group?) and whether this variability can be explained by individual differences factors (i.e., is this individual variability *systematic*?). Even if the overall variability cannot be explained by any particular individual differences, the results will address whether ensembles—that is, subgroups of similarly behaving individuals in terms of their evaluative judgements—are differently influenced by (certain levels of) individual differences (i.e., are the findings at the subgroup level and those of the individuals in said subgroups *mutually inferable*?). In view of the increasing calls for more person-centered approaches to influencing factors in sociolinguistic development (e.g., Ender, 2019; Geeslin et al., 2013; Howard, 2012; Kinginger, 2008; van Compernolle, 2019), these topics warrant much-needed investigation and can inform the variationist SLA landscape as to the dangers of taking group-level estimates at face value without also assessing the learner individuality present in the data. To this end, this study is guided by the following exploratory research questions:

1. To what extent are subject-level sociolinguistic evaluative judgements (i.e., friendliness and intelligence judgements of dialect and standard German, respectively) confluent with group-level patterns in L2 German learners in the Austro-Bavarian context?
2. Is subject-level variability in L2 sociolinguistic evaluative judgements (i.e., the extent to which each participant deviates from the average rating pattern of the group) predicted by individual differences (i.e., systematic differences in standard/dialect proficiency, standard/dialect exposure, and standard/dialect learning motivation)?[4]
3. Which homogeneous ensembles are characteristic of which levels of individual differences variables in adult L2 German learners in Austria? (E.g., are ensembles with high above-average evaluative judgements characteristic of higher proficiency, exposure, and motivation?)

We focus on friendliness and intelligence judgements of dialect and standard German, respectively, as these are common socioindexical attributions for the two varieties in Bavarian-speaking Austria (i.e., dialect speakers are associated with characteristics such as friendliness, sociability etc., whereas standard German speakers are typically judged as more intelligent and educated; Bellamy, 2012; Soukup, 2009).

Methods

Participants

This study includes perception data from 30 participants, all of whom were native English speakers currently living in Bavarian-speaking Austria (Salzburg or Upper Austria; one subject lived in Vienna but worked in Salzburg) with German as an L2. Sixteen subjects were born in the USA, 11 in the United Kingdom, and one in Canada. One participant was born in Peru and one in Japan, both of whom had English-speaking parents and moved to the United States during their childhood. The subject

[4]To be clear, we differentiate between individual *variability* and individual *differences*. Variability is operationally defined as the extent to which each subject deviates from the group-level mean. Individual differences are operationally defined as systematic differences in varietal proficiency, exposure, and motivation.

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

pool was drawn via convenience sampling. Given that the overarching goal of the project was to focus on the credibility intervals of the effects of a range of individual differences variables on sociolinguistic competence rather than homing in on point estimates and binary cutoffs, sample size was not determined via a priori power analyses but rather on the basis of practical considerations (e.g., time, funding, etc.). Our sample pool comprises young and middle-aged adults (M_{age} = 30.03 years, SD = 8.77, range = 20–57). Subjects varied in terms of length of residence (M_{LoR} = 3.88 years, SD = 3.65, range = 0–13.8), self-reported proficiency on a 100-point scale in standard German ($M_{S.G.prof.}$ = 61.0, SD = 23.2, range = 16–100) and Austrian dialect ($M_{dial.prof.}$ = 23.4, SD = 21.0, range = 0–78.8), and self-reported exposure to standard German ($M_{S.G.exp.}$ = 36.55, SD = 26.20, range = 4.9–103) and Austrian dialect ($M_{dial.exp.}$ = 23.95, SD = 24.27, range = 0–92.1), the highest attainable score for each variety being 163.5. See Supplementary Material SF1–SF4 for visualizations of the distributions of the aforementioned variables. The study was approved by the Ethics Committee of the University of Salzburg (EK-GZ 40/2021), and subjects were compensated 20 euro after finishing the experimental procedure in its entirety. For additional information regarding this participant pool (e.g., profession, length of German acquisition, German coursework both in Austria and their homeland), we refer interested readers to the biodata.csv on OSF.

Tasks and procedures

The perception task reported on in this article was one task in a larger test battery, the data collection for which lasted approximately one and a half hours in total. The experimental procedure in the present contribution consisted of a matched-guise task to assess subjects' sociolinguistic evaluative judgements (approximately 7 min) and questionnaires to measure participants' (a) self-reported proficiency in standard German and dialect, (b) self-reported exposure to standard German and dialect, and (c) standard German and dialect learning motivation (approximately 10 min). Data collection took place individually in a quiet and undisturbed room at the participants' convenience.

Matched-guise task (sociolinguistic evaluative judgements)

To assess learners' sociolinguistic evaluative judgements, participants completed a matched-guise task (Lambert et al., 1960). The task targeted subjects' perceptions of Austrian standard German and Austrian dialect varieties. To facilitate comparability, the same voicing stimuli employed in Ender et al. (2017), Ender (2020), and Kaiser et al. (2019) were used. However, due to time constraints, only the four verbal stimuli spoken by the two women speakers were included and the four verbal stimuli spoken by the two men speakers were excluded. The stimuli consisted of everyday greeting sequences: Two greeting sequences were produced in an Austrian standard German variety and two in an Austrian dialect variety. The following orthographically transcribed brief excerpts from the saleswoman guises should illustrate the stark contrasts between varieties[5]:

Standard German: *Ich hab' noch ein paar Semmeln aus der Backstube geholt.*

Austro-Bavarian dialect: *I hob no a poa Semmal aus da Bockstubn ghoit.*

English: *I had to grab a few buns from the bakehouse.*

[5]The full guises, orthographically transcribed, can be found on IRIS.

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

The greeting sequences were tailored to represent different occupations such that one speaker played the role of a bread saleswoman and the other a (woman) doctor (see Kaiser et al., 2019). These speaker occupations were chosen to capture possible differences in "functional prestige" (see Soukup, 2009) of standard German and dialect varieties by L2 speakers. Given that possible differences in functional prestige were of no relevance for the analyses in this contribution, however, the represented occupation of the stimulus speaker was entered as a random intercept in the models to account for possible occupation-specific idiosyncrasies. On each task trial, participants heard four stimulus greeting sequences and were asked to judge the stimuli on the scale focusing on the subjective indexical element of status (question: "How smart is this person?") or solidarity (question: "How friendly is this person?"). Both scales were adopted from Dossey et al. (2020), and cognitive interviews in a pilot study with three participants indicated that subjects associated "smart" and "friendly" with other status- and solidarity-related attributes (e.g., "intelligent" and "nice/likable," respectively). Participants could respond on a 100-point slider scale from "*not at all smart*" to "*very smart*" and "*not at all friendly*" to "*very friendly*," respectively. The presentation of stimuli was blocked by scale following Dossey et al. (2020) such that participants were required to rate each speaker on one scale (e.g., friendliness) before proceeding to the second scale. The order of scale blocks (intelligence vs. friendliness) was randomized, and the order of the four verbal stimuli within each scale block was randomized. In the present contribution, two constellations (Dialect × Friendly; Standard × Intelligent) of judgements are analyzed, as they embody common socioindexical interpretations of the two varieties in Bavarian-speaking Austria (Bellamy, 2012; Kaiser et al., 2019; Soukup, 2009).

Multilingual language profile (varietal exposure and proficiency)

The Bilingual Language Profile (Birdsong et al., 2012) was adapted to create the Multilingual Language Profile questionnaire, which assessed learners' self-reported language history, use, contact, and proficiency with respect to standard German and Austrian dialect. The scores from the modules on language history, use, and contact were used to operationalize the variables standard German and dialect exposure. The original Bilingual Language Profile did not include language contact, but this dimension was included with three items to assess how often subjects hear standard German and dialect with friends, family, and at school/work, as it could well be that subjects hear and come into contact with a variety without necessarily needing to actively use it (e.g., in a lecture, in work-related meetings). Participants could reach a total of 163.5 points per variety. The Multilingual Language Profile also included items on 100-point slider scales exploring subjects' varietal proficiency—that is, proficiency in standard German and dialect with respect to reading, writing, listening, and speaking. The scores were aggregated such that participants could achieve a maximum score of 100 for each variety.

Motivation questionnaire (varietal motivational profiles)

The motivation questionnaire included statement-type items on 100-point slider scales inquiring about learners' peer encouragement for learning dialect (two-item scale: r_{rho} = .65, CI = [0.31, 0.83]) and standard German (two-item scale: r_{rho} = .66, CI = [0.35, 0.87]), interest in dialect (three-item scale: $\alpha = 0.67$) and standard German (three-item scale: $\alpha = 0.62$), and anxiety when speaking dialect (two-item scale: r_{rho} =.78, CI = [0.51, 0.93]) and standard German (two-item scale: r_{rho} = .43, CI = [0.05, 0.73]).

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

Responses for each variety were aggregated and subjects could obtain a maximum score of 100 for each variety. These particular variables were chosen, though broadly speaking, based on qualitative insights gleaned from a similar project conducted in the Swiss-Alemannic context (see, e.g., Ender, 2019) exploring why L2 learners acquire and use standard German and Swiss dialect varieties. Scales reflecting the chosen variables were developed by adapting items from Dörnyei (2010) and Pfenninger and Singleton (2017).

Data analyses

We report two primary statistical analyses, both of which are highly exploratory in nature. First, we analyzed subject-level variability in participants' evaluative judgements by computing Bayesian intercept-only models including by-subject random intercepts to determine a distribution of predicted participant-individual sociolinguistic judgements. Second, using the random effects of the intercept-only models, we used Bayesian multilevel modeling to analyze (a) whether subject-level variability was predicted by individual differences variables (standard/dialect proficiency, exposure, and motivation) and (b) whether individuals within ensembles were more similarly affected by individual differences than was the group.

In the first analysis, two Bayesian multilevel models were fitted using the brms package (Bürkner, 2017) in R (R Core Team, 2020). We modeled friendliness and intelligence evaluative judgements in intercept-only models with by-subject random intercepts (and random intercepts for occupation, see the previous section), the goal being to obtain posterior distributions that reflected each participant's individual evaluative judgement patterns. These posterior distributions served as a measure of subject-level variability—that is, to what extent each individual diverges from the judgement pattern of the group. Because the rating data were necessarily bounded by 0 and 100 by virtue of the slider scale, we opted for the beta distribution (a canonical distribution family for proportion data), which maps the model estimates to the log-odds space using the logit-linking function. Because the beta distribution is bounded to values *between* 0 and 1, the rating data were first divided by 100, and values equal to 0 were manually set to .0001, and values equal to 1 were manually set to .9999. The models included a regularizing, weakly informative prior (Gelman, et al., 2017) for the intercept term, which was normally distributed and centered at 0 with a standard deviation of 5 (in log-odds space)—that is, $Normal(\mu = 0, \sigma = 5)$. All models were fitted with 2,000 iterations (1,000 warm-up). Hamiltonian Monte-Carlo sampling was carried out with four chains in order to draw samples from the posterior predictive distribution.

In the second analysis, we used the posterior of the random effects estimates in the first two intercept-only models to calculate a distribution of 100 plausible predicted values for each participant, reflecting their individual evaluative judgement patterns. Using this as the measure of subject-level variability, we then computed six Bayesian linear mixed-effects models to determine whether subject-level variability was predicted by any individual differences variables. Models included a weakly informative prior—that is, $Normal(\mu = 0, \sigma = 5)$—for all population-level effects and were fitted with 2,000 iterations (1,000 warm-up) and five chains (an extra chain was added in this analysis strictly for computational purposes). We established a region of practical equivalence (ROPE) of $\pm$ 0.08 around a point null value of 0 (in accordance with the suggestions in Kruschke, 2018). For these models, we report mean posterior point estimates for each parameter, along with the 95% highest density interval (HDI; i.e., a

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

type of credible interval, essentially the Bayesian analog to the frequentist confidence interval) and the percentage of the region of the HDI contained within the ROPE. We judge there to be compelling evidence for a given effect when 95% of the HDI of a posterior predictive distribution for a parameter β falls outside the ROPE.

One major advantage of the Bayesian framework—and a primary reason this study makes use of Bayesian models—is its ability to conservatively handle small sample sizes. In essence, the Bayesian framework allows us to investigate the absence of "null effects" (i.e., nonsignificant findings) based on the ROPE. The focus of the analyses is on the distributions of the inquired effects rather than strictly on point estimates and arbitrary significance cutoff points. Whereas frequentist models might generate anticonservative *p* values (i.e., increased Type I error rates) with small samples and underpowered analyses, Bayesian models compute estimates of uncertainty. Therefore, with smaller sample sizes, Bayesian models return estimates with greater uncertainty, which is a more conservative approach. McNeish (2016) makes note in this vein that Bayesian models are particularly suitable for obtaining reliable estimates when dealing with small samples, provided the prior is "set in the vague vicinity of the population value, even with a fairly large variance" (p. 765), thus the justification for the weakly-informative priors mentioned before. The interested reader is referred to (a) McElreath (2015) for more detailed information about the strengths of Bayesian data analysis (as opposed to frequentist methods); (b) Gudmestad et al. (2013) for conceptual advantages of Bayesian analysis in sociolinguistics and SLA; and (c) Franke and Roettger (2019), Garcia (2021), and Vasishth et al. (2018) for tutorials on Bayesian inferential statistics geared toward the language sciences.

Results

The data and code necessary to reproduce the analyses reported in this article are available at https://osf.io/yrqn6/.

Group-level estimates

The first analysis consisted of computing two intercept-only mixed effects models. These provide the group-level estimates of the L2 learners' evaluative judgements of dialect and standard German in terms of friendliness and intelligence, respectively. These two constellations (Dialect × Friendly; Standard × Intelligent) were chosen because they embody common socioindexical interpretations of the two varieties in Bavarian-speaking Austria (Bellamy, 2012; Kaiser et al., 2019; Soukup, 2009). As the numeric model summaries in ST2 and visual model summaries in SF5 in the Online Supplementary Material show, the group-level log odds of rating the dialect as friendly was 1.44 ($\beta = 1.44$, HDI = [−1.54, 4.73])—that is, participants were predicted to rate dialect with a score of approximately 81 on a 100-point scale, though the model predicts plausible scores between 18 and 99. A similar picture emerges when examining participants' evaluative judgements of standard German in terms of intelligence: The interindividual log odds of rating the standard German speakers as intelligent was predicted to be 1.39 ($\beta = 1.39$, HDI = −0.87, 3.55]), or approximately 80 on a 100-point scale, with a credible interval between 30 and 97. These results point toward high interindividual variation in the sociolinguistic evaluative patterns with respect to both the dialect and standard variety. Such extensive variation raises the question as to whether these group-level estimates paint an accurate picture of the individuals in the group.

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

Subject-level variability: Building homogeneous ensembles

To address RQ1—that is, whether subject-level sociolinguistic evaluative judgements of dialect and standard German in terms of friendliness and intelligence, respectively, are confluent with group-level rating patterns—the subject-level random effects of the previous two intercept-only models were analyzed.

Recall that in the Bayesian framework there is no one point estimate for any parameter but rather a whole distribution of plausible values computed via Hamiltonian Monte-Carlo sampling. This holds true for the random effects as well such that there is an entire posterior distribution of individual random effects estimates for each subject rather than a single point estimate. Figure 1 plots posterior distributions of individual intercepts along with the median value (white rhombus) and ± 66% and ± 95% quantile credible intervals (black lines) for each subject. These posterior distributions represent how much each subject differs from the group-level mean estimates (note that zero represents the group-level average estimate of each model). Green shading indicates that the 95% HDI of the subject's posterior distribution did not include zero, providing strong evidence that these subjects deviated starkly *above* the group mean. Blue shading depicts subjects whose 66% HDI did not include zero, thus affording evidence that these subjects also deviated above the group-level average, though to a lesser extent than the green group. The red and gold shading indicate similar tendencies but depict the extent to which each subject deviated *below* the group-level average. Red-shaded subjects' 95% HDI was below the group-level mean (i.e., zero), and gold-shaded subjects' 66% HDI was below the group-level mean.

This subject-centered visual analysis approach allows for two important analyses:

1. We can determine the amount of subject-level variability present in the data by assessing the degree to which subjects deviate from the group-level estimate. This provides us with quantifiable estimates as to how "individually" the present subjects behave in relation to the group.
2. By subsetting participants into groups based on their 95% and 66% HDIs, it is possible to build comparatively homogeneous ensembles in terms of the dimension of sociolinguistic competence under scrutiny (in this case, sociolinguistic evaluative judgements). This allows us to answer myriad calls for further research, particularly

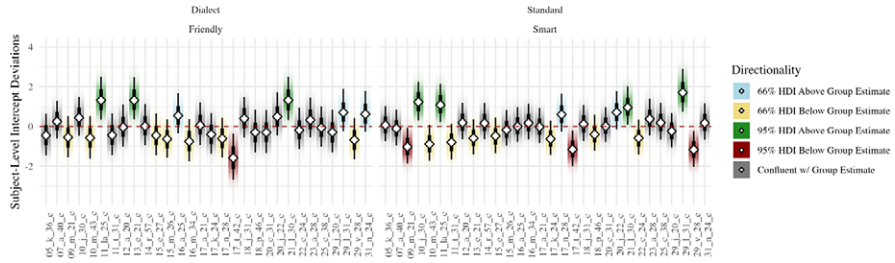

Figure 1. Subject-level variability in L2 evaluative judgements.
Note. The mean evaluative judgements (i.e., the intercepts) for Models 1 (Dialect × Friendly) and 2 (Standard × Smart) are represented at 0 (red dotted line) on the *y*-axis. The plot includes individual posterior point medians (white rhombus), their ± 66% and ± 95% quantile credible intervals (black lines), and the density of the data distribution for each subject. Each gradient interval displays how much subjects deviate from the group rating mean. Red shading indicates that the respective subject's 95% HDI was below the group-level mean and gold shading that the subject's 66% HDI was below the group-level mean. Green shading indicates that the respective subject's 95% HDI was above the group-level mean and blue shading that the subject's 66% HDI was above the group-level mean.

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

with respect to addressing "the extent to which individual L2 learners differ from each other within their corresponding groups [i.e., subgroups within the larger group, MW/SP]" (Geeslin et al., 2013, p. 159). Although other methods have been used to identify (ergodic) ensembles (e.g., [time-series] cluster analysis, see Peng, Jager, & Lowie, 2022; Peng, Lowie, & Jager, 2022), these either (a) are geared toward dense time-series data, as is common in CDST-inspired approaches, and/or (b) do not subset groups *using the group-level mean estimate as a point of reference*. The latter point is particularly important for the current approach such that the overarching goal of the present contribution is to explore whether the group-level estimates paint an accurate picture of the group and how homogeneous ensembles are influenced by individual differences factors *in relation to the group*. Although group-level estimates are often not indicative of the individual, identifying ensembles allows us to form estimates and make predictions specific to smaller subgroups of participants that are mutually inferable with the individuals composing the subgroups (Peng, Lowie, & Jager, 2022). In so doing, we can find patterns that transcend the individual heterogeneity and make inferences more strongly predictive of the trends of certain subgroups of individuals.

Based on these two analyses, we can make the following two observations:

1. **Subject-level variability:** With respect to the subject-level variability, the mean estimate is not representative of many of the individuals. In both models, the group-level estimates would only hold true for approximately half of the subjects (specifically, in Model 1 for 16 and Model 2 for 14 of the subjects), whereas the other half of the participants display exceedingly different amounts of variability, some deviating starkly above and others below the group-level average.
2. **Identifying homogeneous ensembles:** Within the group-level data, there do appear to be clusters of similarly rating individuals. Figure 1 shows these five groups: One is comparatively confluent with the group-level estimates. Two ensembles rate below the group-level average (the red and gold ensembles), whereas the green and blue ensembles rate the respective variety above average. The distinction between the 66% and 95% HDIs stems from the idea that certain groups of individuals deviate more starkly away from the group-level estimates than do others. If we were to explicitly organize subjects according to only one measurement of deviation (e.g., the 95% HDI), this would lead to groups of very homogeneous individuals (those whose 95% HDI is either above or below the group average) and, at the same time, groups of exceedingly heterogeneous individuals, so defeating the purpose of the analysis. By including groups of individuals whose 66% HDI (i.e., approximately 2/3 of their posterior distribution) did not include the group-level estimate, an ensemble can be built of individuals whose sociolinguistic evaluations were still sufficiently different from the group average but to a slightly lesser extent than those whose 95% HDI did not include the group average.[6] As we can see in Figure 1, only a few subjects' 95% HDI did not include zero; the majority of subjects belonged to the gold or blue ensembles, indicating that a large number of subjects deviated from the

[6]We note that any cutoffs for HDIs are entirely arbitrary, as they are descriptive measures of the respective distribution density. With our goal in mind to, in a bottom-up manner, determine an ensemble whose variability measures ranged between the "confluent with group-level estimate" and "extreme deviators," an ensemble for whom approximately 2/3 of their HDI did not include the group-level estimate was deemed appropriate.

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

group-level mean but only few displayed extreme deviations (cf. the green and red ensembles in Figure 1).

Individual differences predictors of subject-level variability

To date, there have been no analyses directly investigating whether subject-level variability can be explained as a function of individual differences. This is a true desideratum, as such analyses can provide a more fine-grained understanding as to the (non)systematic nature of variability in L2 evaluative judgement behavior. In this section and addressing RQ2, subject-level variability is thus subjected to exploratory Bayesian analyses to determine whether the variability in the present subjects' L2 evaluative judgements is explainable as a function of standard/dialect proficiency, exposure, and motivation. Moreover, these analyses address the calls for research in Geeslin et al. (2013) to (a) probe the extent to which variability is linked to linguistic and extralinguistic individual factors and (b) explore individual differences variables beyond proficiency and exposure.

Figure 2 illustrates point estimates and ± 70% and 95% HDI summaries, providing in graphical form an overview of the posterior distributions of the six models assessing whether subject-level variability is explainable as a function of standard German/dialect proficiency, exposure, and motivation (see ST1 in the Supplementary Material for a numerical description). Recall that these six models were computed from the random effects of the previous models in which zero indicated the mean sociolinguistic evaluative judgements of the Dialect × Friendly and Standard × Intelligence models. Based on these group-level analyses of subject-level variability as a function of individual differences, the following three observations can be made:

1. Holding all *z*-scored fixed effects constant at their means (i.e., zero), most intercepts hovered around zero, indicating that the deviations from group-level evaluations were not strongly predicted by the mean-level in the individual differences scrutinized here.
2. Regarding the effects of the individual differences in predicting the subject-level variability in evaluative judgements, there were no credible effects, underscoring that subject-level variability is not necessarily systematic *at the group level* in this sample. However, although not credible, a few select factors did show near-credible predictive power. In particular, higher proficiency in standard German showed a comparatively strong positive directionality in predicting above-average rating patterns, for both the Dialect × Friendly and Standard × Intelligent models. Higher standard German exposure also appeared slightly predictive of above-average rating patterns of dialect in terms of friendliness. Interestingly, whereas higher standard German learning motivation held little to no predictive power, higher dialect learning motivation was comparatively indicative of above-average rating patterns of both models.
3. Finally, the evidence suggests that the few select predictors with the comparatively strongest credibility outlined above were more strongly correlated with above-average rather than below-average ratings. This indicates that the individual differences tended to be slightly predictive of more rather than less native-like patterns of sociolinguistic evaluations (judged against results seen in previous literature on Bavarian-Austrian speakers' sociolinguistic evaluations; see, e.g., Bellamy, 2012;

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

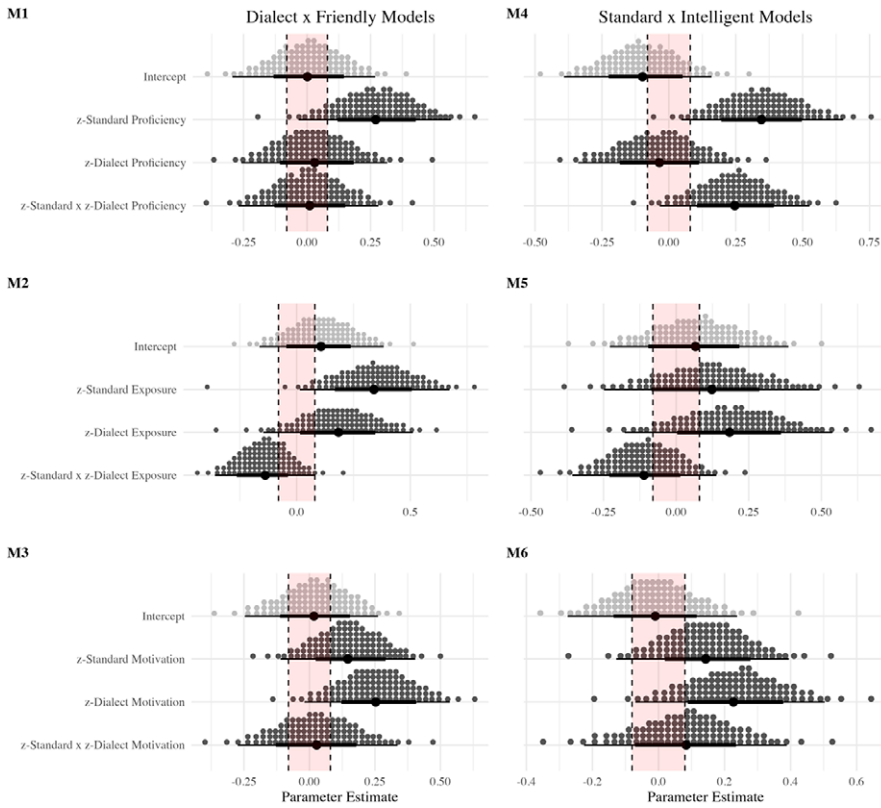

Figure 2. Visual model summaries of Bayesian models showing the effects of individual differences on subject-level variability.

Note. Posterior point estimates and ± 70% and 95% credible intervals for subject-level variability in participants' friendliness ratings of the Austrian dialect (M1, M2, M3) and intelligence ratings of standard German (M4, M5, M6) as a function of standard/dialect proficiency, exposure, and motivation. The light red shading plots the ROPE (i.e., ± 0.08); any posterior distributions whose 95% HDI falls in the ROPE are not considered credible effects.

Kaiser et al., 2019; Soukup, 2009). It must be noted, however, that the group as a whole rated both dialect in terms of friendliness and standard German in terms of intelligence relatively high, implying overall high sociolinguistic competence with respect to mapping social meaning to varietal systems.

Individual differences within and across homogeneous ensembles

In what follows, we address RQ3 concerning whether participants in a respective homogeneous ensemble are characterized by similar levels of individual differences variables. To this end, Figures 3 and 4 display the conditional effects of the predictor variables for the Dialect × Friendly and Standard × Intelligent models, respectively. The homogeneous ensembles discovered in Figure 1 are denoted in this plot via their respective shading (below-average rating: red and gold; above-average rating: green and

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

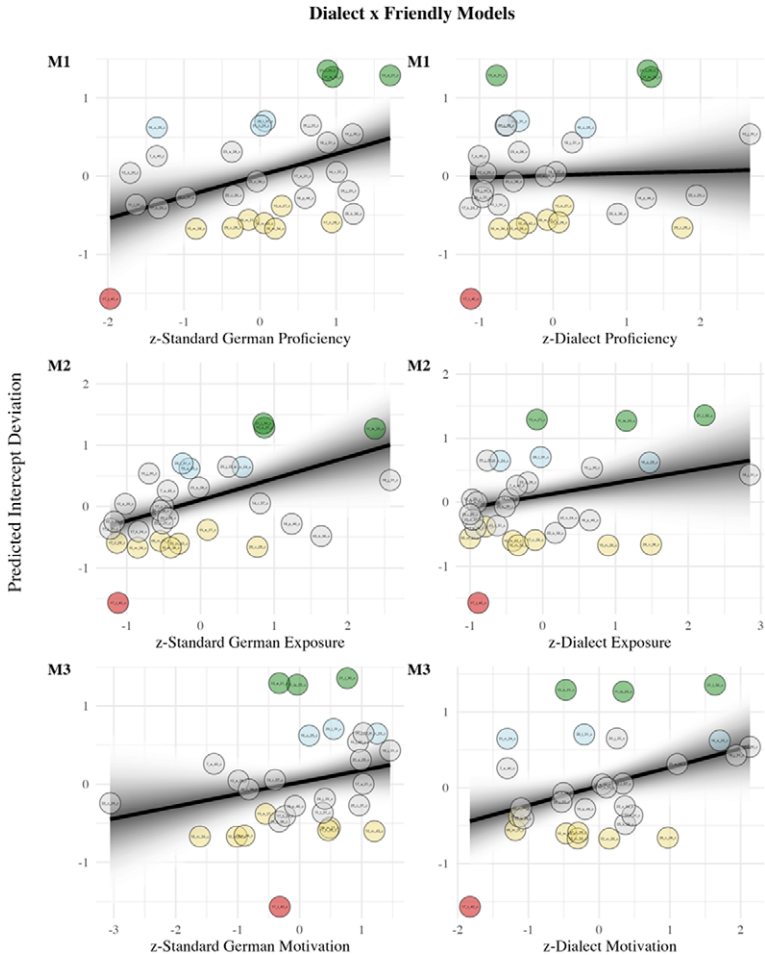

Figure 3. Conditional effects gradient scatter plots for the three Dialect × Friendly models.
Note. The gradient plot displays each subject's mean posterior predicted intercept deviation as a function of the respective *z*-scored predictor. The gray gradient shading around the regression line represents the 95% credible interval, with darker shading indicating more likely values and lighter shading less likely values. The colored points are the ensembles determined via the intercept-only models: Red and gold shading indicate participants whose 95% and 66% HDIs, respectively, are below the group average Dialect × Friendly evaluative judgements; green and blue shading indicate participants whose 95% and 66% HDIs, respectively, are above the group average.

blue). Before we begin, we emphasize that the goal in what follows is *not* to identify which effects are statistically credible for each ensemble, nor do we suggest that visually identified trends are more or less "legitimate" than group-level ones. Rather, by analyzing whether the subjects in a respective ensemble are characterized by similar levels of the individual differences variables, it should be possible to make predictions regarding the effect of certain individual differences that are mutually inferable between the subgroup and individuals therein. This should then allow us to more carefully infer,

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

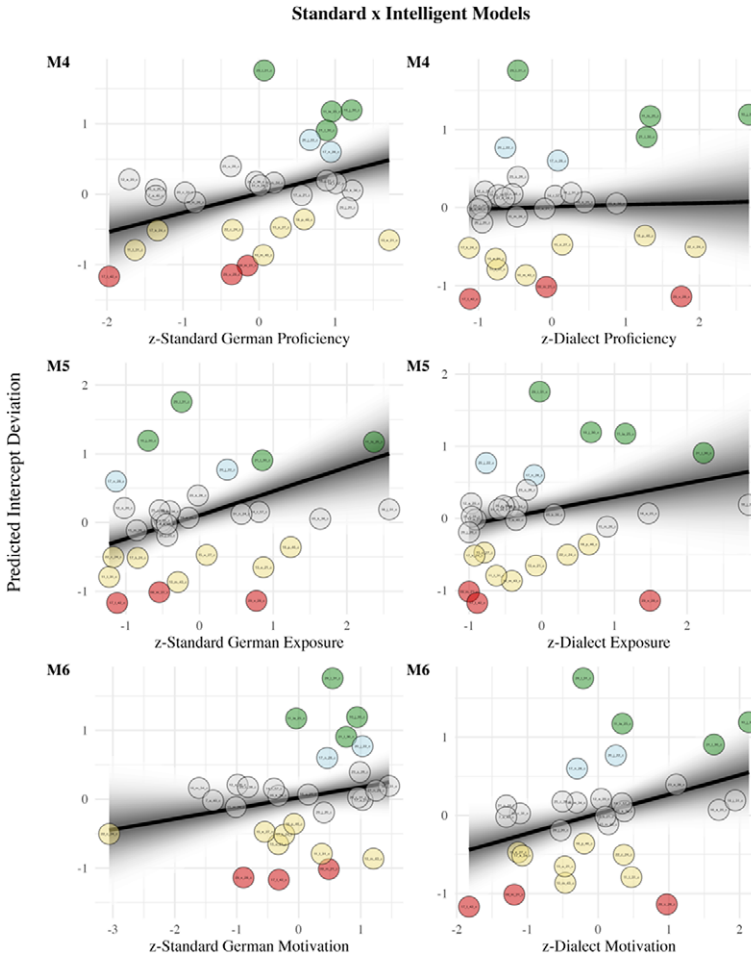

Figure 4. Conditional effects gradient scatter plots for the three Standard × Intelligent models.

or even generalize, to a population of similarly structured individuals as opposed to generalizing mean trends to highly diversified populations.

Based on visual inspection of the effects of the individual differences on the homogeneous ensembles, we can make the following observations:

1. **The Green Ensemble** (95% HDI above average ratings): In the Dialect × Friendly models, the green ensemble ($n = 3$) was characterized by high standard German proficiency, approximately average standard German learning motivation, and comparatively high standard German exposure. Interestingly, there appeared to be a less homogeneous effect of the dialect individual differences predictors, as can be seen in Figure 3. With respect to the Standard × Intelligence models in Figure 4, the green ensemble ($n = 4$) displayed similarities in their high standard German proficiency and slightly above-average standard German motivation. Once again,

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

the effect of dialect individual differences variables was not consistent within the green ensemble.

2. **The Blue Ensemble** (66% HDI above average ratings): Whereas the effects of the individual differences in the Dialect × Friendly model for the blue ensemble ($n = 3$) did not appear homogeneous, the blue ensemble in the Standard × Intelligent models ($n = 2$)—similar to the green ensemble—was characteristic of above-average standard German proficiency and motivation.
3. **The Gold Ensemble** (66% HDI below average ratings): In the Dialect × Friendly models, the gold ensemble ($n = 7$) was particularly characterized by average standard proficiency but below-average amounts of standard German exposure. With respect to the dialect individual differences factors, the gold ensemble displayed comparatively average dialect proficiency but was widespread with respect to dialect exposure and motivation. In the Standard × Intelligence models, the ensemble ($n = 7$) appeared particularly characterized by average standard German and dialect learning motivation.
4. **The Red Ensemble** (95% HDI below average ratings): In the Dialect × Friendly model, only one subject's 95% HDI deviated below the group average, and this subject displayed below-average levels in the standard German and dialect-related individual differences measures. Given that there was only one subject in the red ensemble in Model 1, however, future research with a larger sample is needed to determine the degree of systematicity concerning lower levels of individual differences in varietal proficiency, exposure, and motivation and ensembles deviating starkly below group-level means in Dialect × Friendly judgements. In the Standard × intelligent models, however, the red ensemble ($n = 3$) was not characterized by any particular level of the individual differences.
5. **The Gray Ensemble** (confluent with the group-level estimates): Interestingly, this ensemble showed the least amount of homogeneity with respect to the effects of the individual differences factors. Visual inspection did, however, reveal that in both models, the gray ensemble (Model 1: $n = 16$; Model 2: $n = 14$) appeared particularly characteristic of slightly below-average dialect proficiency and exposure.

Importantly, by exploring the influence of individual differences on ensembles rather than the group as a whole, we can identify predictive effects that would have been neglected or perhaps even misinterpreted if only addressing group-level model estimates. For example, the group-level estimate regarding subject-level variability as a function of standard proficiency in the Dialect × Friendly model suggested (though did not reach credibility) that with higher standard proficiency subjects deviate higher above group-average evaluative judgements. However, as the conditional effects plot shows, the effect of proficiency is only (comparatively) homogeneously predictive for the green and gold ensembles. Given that there was only one participant in the red ensemble in Model 1 (Dialect × Friendly judgements), we could not determine the systematicity of individual differences levels, but we do note that the blue and gray ensembles show no similar effect of standard proficiency on their subject-level variability. A similar picture emerges with respect to, for example, the effect of standard German exposure in the Dialect × Friendly model and standard proficiency in the Standard × Intelligence model. Thus, taking the group-level effects at face value would have produced findings that obscure individual learner and ensemble diversity and may have led us to conclude no credible effects. However, the ensemble-guided approach outlined here underscores that certain individual differences only appear predictive at certain levels and for certain ensembles, a finding that may have been neglected if

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

strictly focusing on group-level results. The advantage of this approach is as follows: By identifying *which* ensembles at *which* levels of individual differences are more homogeneous, the subgroup is mutually inferable with the individuals therein, which should then allow us to make more careful predictions and generalizations of individual cases (Peng, Jager, & Lowie, 2022). In so doing, we can avoid carelessly attempting to generalize mean scores to much larger and more diversified populations.

Discussion

One of the underlying goals in variationist SLA is to determine how L2 learners acquire target-like variation (both in production and perception) and which (set of) factors influences this (Kanwit, 2022). However, the acquisition of variation is an individually owned process (Ender, 2019; Howard, 2012; Kinginger 2008), which calls into question whether group-level estimates based on quantitative analytical models can provide as insightful results into the acquisition of variation as once hoped (Geeslin et al., 2013). This calls for person- and ensemble-centered analyses, such as those presented in this contribution with respect to L2 German learners in the Austro-Bavarian naturalistic sphere. Specifically, using the proposed ensemble-centered analyses, we can compare group-level estimates of L2 evaluative judgements of standard German and dialect varieties with those of the individual so as to determine which effects of, for example, individual differences are characteristic for certain individuals or groups of similarly behaving individuals (i.e., homogeneous ensembles).

We ran two group-level analyses: The first addressed RQ1 and consisted of computing two intercept-only mixed effects models (a) to obtain the mean-level estimate of subjects' sociolinguistic evaluative judgements of dialect in terms of friendliness and standard German in terms of intelligence and (b) to use the random effects to obtain measures of how much each individual deviated from the group-level mean (i.e., subject-level variability). This analysis showed extreme variation both above and below the group-level means of the models, indicating that the mean estimates did not paint an accurate picture of the group. In a follow-up analysis, in line with RQ2, we assessed whether this subject-level variability could be predicted by individual differences variables (standard German/dialect proficiency, exposure, and motivation). The models found no credible effects for any of the predictors. Using the intercept-only models and measures of deviation, however, it was possible to statistically capture subgroups of similarly rating individuals, which we used to operationally define homogeneous ensembles. By creating these homogeneous ensembles, the goal was to find patterns that transcend the heterogeneity of the individual so as to make inferences more strongly predictive of the trends of certain subgroups of individuals (see Molenaar, 2015; Molenaar & Campbell, 2009). This ensemble-centered visual analysis approach, indeed, allowed us to identify predictive effects for certain ensembles that would have perhaps been neglected if only addressing group-level model estimates (RQ3).

The results of this study both conflict with and underscore previous results, although all comparisons must be taken with a careful grain of salt given differences in production versus perception and the respective variable(s) under scrutiny. On the one hand, our findings stand in stark contrast with Regan (2004), who found in her production data that her small homogeneous group of learners followed similar paths in acquiring the French sociolinguistic variable *ne*. Our perception-based results—another dimension of sociolinguistic competence—display excessive subject-level variability, pointing toward heterogeneity in sociolinguistic development. This confirms

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

Geeslin et al.'s (2013) note that sociolinguistic competence operationalized using different linguistic structures and collected using other tasks imposes different amounts of subject-level variability. On the other hand, Geeslin et al. (2013) found that variability in the acquisition of variation tended to decrease with increasing proficiency levels. As can be seen in Figures 3 and 4, the green ensemble was characterized by a very homogeneous effect of standard German proficiency, indicating that degrees of subject-level variability were more similar for learners with high standard proficiency; we must note, however, that this homogeneous effect did not carry over to *dialect* proficiency. This finding was particularly interesting, given that Ender (2020) found dialect proficiency (using a translation task) to be a strong predictor of higher evaluative judgements toward dialect. Our results indicating a negligible role of dialect proficiency, indeed, contradict hers; however, there is likely a methodological rationale to this: For one, Ender (2020) used a global rating scale and did not differentiate between indexical domains. For another, her sample comprised both L1 and L2 speakers. The divergent results may thus simply be in function of methodological and sample-related discrepancies. In any case, these results do not dispute hers, nor can her results fully generalize to this sample pool. These inconsistent findings do, however, incite the need for further investigations concerning the role of, for example, varietal proficiency on L2 evaluative judgements.

The results of this study should inform future variationist SLA (and fields using quantitative methods as a whole) that variability (a) should not be neglected or written off as white noise and (b) is not necessarily coincidental but might underlie some form of systematicity. Methodologically, using random effects to address the systematicity of variability presents an innovative way to explore the degree of individuality in the acquisition of sociolinguistic competence in any given sample (see also Drager & Hay, 2012), as it allows us to compare and contrast group-level estimates with estimates of the individual. Such an approach does justice to the call in Geeslin et al. (2013) not to "obscure individual differences through the reporting of only group norms of use" (p. 156).

In terms of limitations, we acknowledge the fact that the sample size used to propose this methodological innovation is quite small. Future research employing (similar) person-centered analysis methods should (a) increase both the number of stimuli and participants and (b) complement the ensemble-level analyses with qualitative data that might aid in explaining the individual- and ensemble-related trends found. Conceptually speaking, it must also be noted that the generalizability of our findings may be restricted by the cross-sectional data set. Although we could identify homogeneous ensembles, the lack of intensive time-serial data precludes us from both identifying and affirming ergodic ensembles, given that we could not assess the stationarity criterion of the ergodic theorem (i.e., when the group and individual mean and variance remain consistent over time). Therefore, our approach only allowed us to assess whether analyses of interindividual variation translate to the *individuals* but not to their *intra*individual variation (i.e., variation across time). As mentioned, dense serial measures would prove opportune for this, but the approaches demonstrated in the present contribution could be (adapted and) applied. Even in light of this conceptual limitation, the question as to whether group-level effects translate to the individual is not equivocal: As we have argued, this ensemble approach presents an innovative way to address the *group-to-individual generalizability* issue in cross-sectional data and transcend individual variability so as to make careful and more accurate generalizations of individual cases to wider populations of similarly behaving L2 German learners in the Austro-Bavarian naturalistic sphere.

https://doi.org/10.1017/S0272263123000177 Published online by Cambridge University Press

Conclusion

The present study is the first to investigate subject-level variability in L2 sociolinguistic evaluative judgements and explore whether this variability is, to any extent, systematic and characterizable as a function of individual differences in standard German/dialect proficiency, exposure, and motivation. The results suggest that group-level estimates rarely paint an accurate picture of the group. It is thus necessary to innovatively tackle the challenge of developing methods for integrating quantitative, population-level estimates with person- and/or ensemble-centered approaches so as not to lose sight of the individual among increasingly sophisticated statistical practices. Random effects present one meaningful way to keep track of the individual, which can be used to analyze the systematicity of subject-level variability. In this sample, we found that subject-level variability in sociolinguistic evaluative judgements as an outcome variable was not credibly predictable by global individual differences in standard German/dialect proficiency, exposure, and motivation. By building homogeneous ensembles (i.e., subgroups of similarly behaving individuals with respect to their evaluative judgements), however, it was possible to assess whether certain ensembles were characteristic of certain levels of the individual differences variables under scrutiny. The results showed that, although not every ensemble was affected similarly by an individual differences factor in every measure, the subject-level variability of individuals within their respective ensembles appeared to underlie systematicity, although to different extents across models. This goes to show that it would be irrational to write off variability as immaterial when applying quantitative analyses; rather, subject-level variability should (at the very least) be reported or, better yet, undergo its own systematic analysis before assuming that the population-level effects paint an accurate picture of the group. Moreover, the homogeneous-ensemble approach used here could be expanded to aid in more closely assessing which effects are predictive for which subgroups of individuals. For example, future models could use the random effects directly from the regression models to identify ensembles. Building from this, instead of presenting strictly population-level effect sizes (and perhaps accompanying random effects), future analyses could combine qualitative data with the homogeneous ensembles identified via the random effects so as to better rationalize how and why certain ensembles are affected in a certain way, the outcome being inferences and generalizations more strongly predictive of the trends of certain subgroups and the individuals therein.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263123000177.

Data availability statement. The experiment in this article earned Open Data and Open Materials badges for transparent practices. The materials are available at https://osf.io/yrqn6/; https://www.iris-database.org/details/jDmnX-zUH43.

Acknowledgments. We are greatly indebted to the participants for their enthusiastic participation and patience during the data collection. We further wish to thank the anonymous reviewers for their incredibly helpful comments and suggestions. Any remaining errors remain our own.

Funding. This article was funded by *Salzburg Stadt: Kultur, Bildung und Wissen*, which is hereby gratefully acknowledged.

Competing interest. The authors declare no competing interests.

References

Bayley, R., & Langman, J. (2004). Variation in the group and the individual: Evidence from second language acquisition. *International Review of Applied Linguistics in Language Teaching*, *42*, 303–318.

Bellamy, J. (2012). *Language attitudes in England and Austria: A sociolinguistic investigation into perceptions of high and low-prestige varieties in Manchester and Vienna*. Franz Steiner Verlag.

Birdsong, D., Gertken, L., & Amengual, M. (2012). *Bilingual Language Profile: An easy-to-use instrument to assess bilingualism*. University of Texas. https://sites.la.utexas.edu/bilingual/

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Bülow, L., & Pfenninger, S. E. (2021). Introduction: Reconciling approaches to intra-individual variation in psycholinguistics and variationist sociolinguistics. *Linguistics Vanguard*, *7*, Article 20200027. https://doi.org/10.1515/lingvan-2020-0027

Chappell, W., & Kanwit, M. (2022). Do learners connect sociophonetic variation with regional and social characteristics? The case of L2 perception of Spanish aspiration. *Studies in Second Language Acquisition*, *44*, 185–209.

Clark, L., & Schleef, E. (2010). The acquisition of sociolinguistic evaluations among Polish-born adolescents learning English: Evidence from perception. *Language Awareness*, *19*, 299–322.

Davydova, J., Tytus, A., & Schleef, E. (2017). Acquisition of sociolinguistic awareness by German learners of English: A study in perceptions of quotative be like. *Linguistics*, *55*, 783–812.

Dörnyei, Z. (2006). Individual differences in second language acquisition. In K. Bardovi-Harlig & Z. Dörnyei (Eds.), *Themes in SLA research* (pp. 42–68). John Benjamins.

Dörnyei, Z. (2010). *Questionnaires in second language research* (2nd ed.). Routledge.

Dossey, E., Clopper, C. G., & Wagner, L. (2020). The development of sociolinguistic competence across the lifespan: Three domains of regional dialect perception. *Language Learning and Development*, *16*, 330–350.

Drager, K., & Hay, J. (2012). Exploiting random intercepts: Two case studies in sociophonetics. *Language Variation and Change*, *24*, 59–78. https://doi.org/10.1017/S0954394512000014

Ellis, R. (2004). Individual differences in second language learning. In A. Davies & C. Elders (Eds.), *The handbook of applied linguistics* (pp. 525–551). Blackwell Publishing.

Ender, A. (2019). *Dialekt-standard-variation im ungesteuerten zweitspracherwerb des Deutschen. Eine soziolinguistische analyse zum erwerb von variation bei erwachsenen lernenden* [Unveröffentlichte Habilitationsschrift]. Universität Freiburg.

Ender, A. (2020). Zum zusammenhang von dialektkompetenz und dialektbewertung in erst- und zweitsprache. In M. Hundt, A. Kleene, A. Plewnia, & V. Sauer (Eds.), *Regiolekte. Objektive sprachdaten und subjektive sprachwahrnehmung* (pp. 77–102). Narr Francke Attempto Verlag.

Ender, A., Kasberger, G., & Kaiser, I. (2017). Wahrnehmung und bewertung von dialekt und standard durch jugendliche mit Deutsch als erst- und zweitsprache. *ÖDaF-Mitteilungen*, *33*, 97–110.

Franke, M., & Roettger, T. (2019). *Bayesian regression modeling (for factorial designs): A tutorial.* https://github.com/michael-franke/bayes_mixed_regression_tutorial/blob/master/text/bmr_tutorial.tex

Garcia, G. D. (2021). *Data visualization and analysis in second language research*. Routledge.

Geeslin, K. L., Linford, B., Fafulas, S., Long, A., & Díaz-Campos, M. (2013). The L2 development of subject form variation in Spanish: The individual vs. the group. In J. Cabrelli, G. Lord, A. De Prada Pérez, & J. Aaron (Eds.), *Selected proceedings of the 16th Hispanic linguistics symposium* (pp. 156–174). Cascadilla Proceedings Project.

Geeslin, K. L., & Long, A. Y. (2014). *Sociolinguistics and second language acquisition. learning to use language in context*. Routledge.

Geeslin, K. L., & Schmidt, L. B. (2018). Study abroad and L2 learner attitudes. In C. Sanz & A. Morales-Front (Eds.), *The Routledge handbook of study abroad research and practice* (pp. 387–405). Routledge.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*, 1–13.

George, A. (2014). Study abroad in central Spain: The development of regional phonological features. *Foreign Language Annals*, *47*, 97–114.

Gudmestad, A., Edmonds, A., & Metzger, T. (2020). Modeling variability in number marking in additional-language Spanish. *Journal of the European Second Language Association*, *4*, 24–34. https://doi.org/10.22599/jesla.67

Gudmestad, A., House, L., & Geeslin, K. L. (2013). What a Bayesian analysis can do for SLA: New tools for the sociolinguistic study of subject expression in L2 Spanish. *Language Learning*, *63*, 371–399.

Guy, G. (1980). Variation in the group and in the individual: The case of final stop deletion. In W. Labov (Ed.), *Locating language in time and space* (pp. 1–36). Academic Press.

Hiver, P., Al-Hoorie, A. H., & Evans, R. (2022). Complex dynamic systems theory in language learning: A scoping review of 25 years of research. *Studies in Second Language Acquisition*, *44*, 913–941.

Howard, M. (2012). The advanced learner's sociolinguistic profile: On issues of individual differences, second language exposure conditions, and type of sociolinguistic variable. *The Modern Language Journal*, *96*, 20–33.

Howard, M., Mougeon, R., & Dewaele, J.-M. (2013). Sociolinguistics and second language acquisition. In R. Bayley, R. Cameron, & C. Lucas (Eds.), *The Oxford handbook of sociolinguistics* (pp. 340–359). Oxford University Press.

Kaiser, I., Ender, A., & Kasberger, G. (2019). Varietäten des österreichischen Deutsch aus der hörerinnenperspektive: Diskriminationsfähigkeiten und sozio-indexikalische interpretation. In L. Bülow, A. Fischer, & K. Herbert (Eds.), *Dimensionen des sprachlichen raums: Variation—mehrsprachigkeit—konzeptualisierung* (pp. 341–362). Peter Lang Verlag.

Kanwit, M. (2017). What we gain by combining variationist and concept-oriented approaches: The case of acquiring Spanish future-time expression. *Language Learning*, *67*, 461–498. https://doi.org/10.1111/lang.12234

Kanwit, M. (2019). Beyond the present indicative: lexical futures as indicators of development in L2 Spanish. *The Modern Language Journal*, *103*, 481–501. https://doi.org/10.1111/modl.12566

Kanwit, M. (2022). Sociolinguistic competence: What we know so far and where we're heading. In K. L. Geeslin (Ed.), *The Routledge handbook of second language acquisition and sociolinguistics* (pp. 30–44). Routledge.

Kanwit, M., & Geeslin, K. L. (2020). Sociolinguistic competence and interpreting variable structures in a second language: A study of the copula contrast in native and second-language Spanish. *Studies in Second Language Acquisition*, *42*, 775–799. https://doi.org/10.1017/S0272263119000718

Kinginger, C. (2008). Language learning in study abroad: Case studies of Americans in France. *The Modern Language Journal*, *92*, 1–131.

Kliesch, M., & Pfenninger, S. E. (2021). Cognitive and socioaffective predictors of L2 microdevelopment in late adulthood: A longitudinal intervention study. *The Modern Language Journal*, *105*, 237–266.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*, 270–280.

Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford University Press.

Lambert, W., Hodgson, R., Gardner, R., & Fillenbaum, S. (1960). Evaluational reactions to spoken language. *Journal of Abnormal and Social Psychology*, *60*, 44–51.

Lowie, W. M., & Verspoor, M. H. (2019). Individual differences and the ergodicity problem. *Language Learning*, *69*, 184–206.

Lyster, R. (1994). The effect of functional-analytic teaching on aspects of French immersion students'sociolinguistic competence. *Applied Linguistics*, *15*, 263–287.

McElreath, R. (2015). *Statistical rethinking. A Bayesian course with examples in R and Stan*. Chapman Hall.

McKenzie, R. M. (2008). Social factors and non-native attitudes towards varieties of spoken English: A Japanese case study. *International Journal of Applied Linguistics*, *18*, 63–88.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 750–773.

Molenaar, P. C. M. (2015). On the relation between person-oriented and subject-specific approaches. *Journal for Person-Oriented Research*, *1*, 34–41.

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*, 112–117.

Peng, H., Jager, S., & Lowie, W. M. (2022). A person-centred approach to L2 learners' informal mobile language learning. *Computer Assisted Language Learning*, *35*, 2148–2169.

Peng, H., Lowie, W., & Jager, S. (2022). Unravelling the idiosyncrasy and commonality in L2 developmental processes: A time-series clustering methodology. *Applied Linguistics*, *43*, 891–911. https://doi.org/10.1093/applin/amac011

Pfenninger, S. E., & Singleton, D. (2017). *Beyond age effects in instructional L2 learning. revisiting the age factor*. Multilingual Matters.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Regan, V. (2004). The relationship between the group and the individual and the acquisition of native speaker variation patterns: A preliminary study. *IRAL—International Review of Applied Linguistics in Language Teaching*, *42*, 335–348.

Regan, V. (2010). Sociolinguistic competence, variation patterns and identity construction in L2 and multilingual speakers. *EUROSLA Yearbook*, *10*, 21–37.

Regan, V., Howard, M., & Lemée, I. (2009). *The acquisition of sociolinguistic competence in a study abroad context*. Multilingual Matters.

Schmidt, L. B., & Geeslin, K. L. (2022). Developing language attitudes in a second language. Learner perceptions of regional varieties of Spanish. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, *35*, 206–235.

Soukup, B. (2009). *Dialect use as interaction strategy. A sociolinguistic study of contextualization, speech perception, and language attitudes in Austria*. Braumüller.

Tagliamonte, S. (2006). *Analysing sociolinguistic variation*. Cambridge University Press.

van Compernolle, R. A. (2019). Constructing a second language sociolinguistic repertoire: A sociocultural usage-based perspective. *Applied Linguistics*, *40*, 871–893. https://doi.org/10.1093/applin/amy033

van Compernolle, R. A., & Williams, L. (2012). Reconceptualizing sociolinguistic competence as mediated action: Identity, meaning-making, agency. *The Modern Language Journal*, *96*, 234–250.

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161.

Cite this article: Wirtz, M. A. and Pfenninger, S. E. (2023). Variability and individual differences in L2 sociolinguistic evaluations: The GROUP, the INDIVIDUAL and the HOMOGENEOUS ENSEMBLE. *Studies in Second Language Acquisition,* 45: 1186–1209. https://doi.org/10.1017/S0272263123000177