LingBaW. Linguistics Beyond and Within 10 (2024), 7–32

Pushing boundaries in the measurement of language attitudes: Enhancing research practices with the L'ART Research Assistant app

Florian Breit[1,2], Marco Tamburelli[1], Ianto Gruffydd[1,3] and Lissander Brasca[1]
[1]Bangor University; [2]University College London; [3]Cardiff University, UK

Abstract

The importance of methodological developments has recently been emphasised both in language attitude research specifically (Kircher & Zipp 2022), and across linguistics and the social sciences more broadly, where there has been a particular focus on replicability (Sönnig & Werner 2021; Kobrock & Roettger 2023). One aspect of this concerns the adoption of more open, consistent, and comparable implementations of method.

We introduce a new digital application (the L'ART Research Assistant) for research in multilingualism and language attitudes. Designed specifically for work with populations speaking a majority and a regional/minority/minoritised/heritage language, the app implements reference versions of some common research methods and tasks. This benefits the research community by enhancing consistency and comparability within and across studies and by improving replicability and reproducibility.

We discuss technical and methodological considerations behind the app and illustrate its use with a brief case study of language attitudes across three European communities whose regional/minority languages receive radically different degrees of socio-political recognition: Lombard (Italy), Moselle-Franconian (Belgium), and Welsh (UK). The case study demonstrates not only how the app facilitates research across different communities that is easily comparable, results also reveal fundamental differences in attitude scores depending on the methods employed (AToL v. MGT). Consequently, we argue that there is a need to move toward both the adoption of more consistent, comparable methods as well as toward a more holistic approach to measuring language attitudes, where a battery of tests — as opposed to a single measure — should become the norm.

Keywords: language attitudes; bilingualism; methodology; research tools; replicability

1. Introduction

Speaker attitudes are considered a fundamental barometer for the current and future vitality of a language, with recent work emphasising the importance of methodological developments (Kircher & Zipp 2022). This, together with the growing concern surrounding the replicability of results across the social sciences, including linguistics (e.g. Sönnig & Werner 2021; Grieve

LingBaW. Linguistics Beyond and Within, e-ISSN: 2450-5188
DOI: 10.31743/lingbaw.18005

2021; Kobrock & Roettger 2023), calls for urgent developments in research practices, including the adoption of more consistent and comparable implementations of methods. To this end, we present the L'ART Research Assistant, a newly developed, freely available open-source application for the collection, storage and transfer of data for research in bilingualism and language attitudes, with a particular focus on bilingual populations who speak a majority language (ML) and a regional/minority/minoritised or heritage language (RML). The app aims to make research in bilingualism — and particularly on language attitudes — easier, more comparable, and readily replicable. The most recent release of the app as of writing (version 0.5.2) implements a digital informed consent facility, a comprehensive background questionnaire in the form of the Language and Social Background Questionnaire (Anderson et al. 2018), the Attitudes Towards Languages Scale (Schoel et al., 2013), the Verbal Guise/Matched Guise Technique (Lambert et al. 1960; and e.g. Markel et al. 1967) and a simple memory game which can be employed as a distractor task.

In this article, we first discuss the general advantages provided by apps like the L'ART Research Assistant, before describing the currently implemented research tools and their respective adaptations. We then illustrate the practical use of the L'ART Research Assistant by briefly discussing results from a case study on linguistic attitudes carried out using the app, which looked at three bilingual communities: Lombard—Italian in Northern Italy, Moselle-Franconian—German in East Belgium and Welsh—English in North Wales.

2. The L'ART Research Assistant: main advantages

Integrated digital research toolkits such as the L'ART Research Assistant provide many advantages to researchers over paper-and-pencil or word-processor based questionnaires, as well as questionnaires implemented through online survey platforms, especially where use of a single tool finds broader adoption across several studies, populations, and/or research groups. The advantages of adopting such toolkits include:

- less work for the researcher, who can largely rely on pre-implemented logic and only needs to adapt stimuli where no suitable ones are yet provided for the target population;
- enhanced consistency and comparability within and across studies, since — apart from translation/localisation-related content — the presentation, data types, validation, coding and output formats remain consistent across use instances;
- improved transparency and replicability/reproducibility, because the entire source code is openly available and version controlled, so that referencing specific versions of the app/tasks allows other researchers to view and reconstruct tasks as they were administered at the time the research was carried out.

Note that while we also mention online survey platforms here, the primary focus of these platforms is different from that of our app: the L'ART Research Assistant is (at least presently) aimed foremost at research conducted offline, whether in a lab or in the field. Particularly where fieldwork on regional and minority languages is concerned, the research setting often

only provides limited or unstable internet connection, if internet connection is available at all — a significant factor in our decision to adopt an offline-first approach.[1]

2.1. Input validation and consistency

Paper/word-processor based questionnaires are especially prone to human error. Participants may omit questions, forget to enter some information, or provide data that is inconsistent in some manner. Researchers may accidentally omit questions or parts thereof, or alter layout and formatting in an unintended manner that could influence responses. The same document opened in a different word processor or on a different machine might produce different layouts and/or affect functionality such as built-in form fields. Some participants may look ahead or back, and alter their answers depending on what they see next, while others may decide not to do so, leading to a loss of control and consistency concerning the data collection flow. Finally, in implementing or adapting the questionnaire, researchers may accidentally omit questions or parts thereof, or alter layout and formatting in an unintended manner that could influence responses.

Questionnaires implemented using common online survey platforms (e.g. Qualtrics, SurveyMonkey) usually succeed in reducing at least some of these error sources e.g. through marking certain fields required and through basic input validation (e.g. a 'date' data field will only accept a valid date). Many platforms also have features that allow researchers to implement at least some conditional logic, though experience shows that the degree to which individual researchers take advantage of such advanced features varies greatly. What they cannot address are those aspects that depend principally on the researcher(s) implementing the questionnaire, as well as to some degree the flow and layout, which may be limited by the specific platform chosen.

A purposely developed app also cannot entirely eliminate all such sources of errors and inconsistencies. It can, however, considerably reduce their occurrences (e.g. Vergnaud et al. 2011) and ensure much greater consistency in that endeavour compared to situations where we have many different more-or-less independent implementations of what is supposedly the same instrument (questionnaire or other research task). The separation of control flow and validation logic from stimuli and text prompts/translations makes it possible to focus efforts in a way that leads to richer and more consistent validation and control flow throughout, which then is not dependent on whether the individual researcher has the time and/or skills to implement this logic themselves.

In the L'ART Research Assistant, not only must participants complete all required questions before they are able to continue, but responses are also validated both in the user interface and in the underlying data models, so that e.g. an invalid date (e.g. 31st February 1983) or a repeated entry of the same language (e.g. a participant who reports that they speak

[1] Many choices in the design and implementation of the software were specifically made such that the software could potentially be extended to offer both offline and online versions of its research tasks in the future, though we currently do not have the resources available to pursue this further.

Spanish as both their second and fourth language) will be rejected and the user is provided automatic feedback on the error.

In many cases, the user interface design already pre-empts certain errors or inconsistent responses. For example: a question on whether a participant uses corrective lenses is only presented if they have indicated impaired eyesight; sliders (some with an adjacent "not applicable" option) are used where a range value (e.g. 0-100) is to be collected; when a bilingual user is prompted to state the languages they speak, two blank language fields are presented as required initially, with location-based auto-suggestion of language options as they complete the field, ensuring that bilingual participants minimally enter at least two languages (participants can add additional languages);[2] further, these questions automatically generate relevant follow-on questions specifically referring back by name to the languages participants entered earlier, ensuring that participants always provide complete information on all their languages.

Potential confounds introduced by participants looking ahead or returning to adjust answers at a later point in the task are controlled by disabling facilities to return to previous screens and only allowing participants to advance to the next screen once the current screen has been fully and validly completed. For example, in the AToL-C, participants cannot return to change their ratings for Welsh once the task has advanced to solicit ratings on English. Similarly, in the audio guise task participants can only listen to a guise once, and they are unable to return and change their ratings for a previous guise once they have moved on. This makes the data collection procedure more consistent and rigorous by minimising presentation-based cross-participant and cross-study confounds.[3]

Another common source of human error is introduced where paper or word-processor based questionnaire data is transferred to a digital format suitable for spreadsheet and statistical applications. Applications such as the L'ART Research Assistant entirely sidestep this by recording data in a widely used data exchange format (JSON), which is easily exported/imported, shared, and inspected, as well as being compatible with most tools in the modern data-pipeline. Note also on this point again the advantage of consistency in the specific layout of the data files where a single tool is adopted (whether that be our app or some other tool): even though for example both survey platforms A and B might offer the export of responses as JSON files, it is unlikely that these JSON files will follow the same, consistent format across platforms, potentially requiring extensive checks and pre-processing to combine data from different sources.

---

[2] It is, in the authors' experience, not unusual to find participant background data (usually from paper/word-processor based questionnaires) with conflicting information, which would suggest, for example, that some multilinguals only speak a single language.

[3] Researchers can temporarily unlock the facility to return to a previous screen, e.g. in cases where an error occurred. Requiring manual intervention ensures that researchers will be aware of such issues and can keep appropriate records.

2.2. Data security and shareability

Particular attention was paid to designing the app and its tools such that no identifying sensitive personal information needs to be recorded. Data is referenced to participants only by their participant ID. Participant IDs must be pre-assigned by the researcher, meaning that sensitive participant information will be stored and managed entirely separately from the research data collected with the app, as is best practice for data protection and security.

Data files from participant responses are stored as individual JSON files, each identified by both the participant ID and an automatically generated Universally Unique Identifier (UUID). The UUID reduces chances of naming clashes with duplicate IDs or multiple responses from the same participant for the same task (e.g. due to repeated testing), even when collating and merging large data sets from different studies. With this system, the chance of duplication or accidental overwrites due to filename conflicts is practically nil even when large data sets are simply pasted together in place.[4]

Data backup is also facilitated by the increased unicity of response data files — researchers need not worry about potential clashes when copying large sets of collated data files for backup, and since they are stored as plain files in the systems' user profile location (specifically, the roaming profile on systems supporting this), they are easily captured by many standard solutions for the synchronisation and backup of user profile data which may already be institutionally deployed by system administrators, in addition to being easily targetable for cloud synching and backup with user-level tools such as rsync[5]. An additional, integrated cloud-based backup and synchronisation option is planned for future versions of the app.

Shareability of data is also facilitated by this approach: researchers can simply transfer/upload/share files for all relevant or selected responses as they would any other file. Being text-based JSON files, they can be easily compressed, transferred, and validated, and are compatible with a wide range of data tools and programming packages. The underlying implementations of the app's data models are currently being ported to Pydantic[6], one of the most widely used Python packages for data models and validation, which will further facilitate the interchange and use of research data, for instance by making the models more easily accessible programmatically for Python users, and by providing auto-generated JSON Schemas for the files produced by each research task, so that shared or imported files can be easily revalidated on import.

---

[4] Statistically, we would have to generate one data file per second for approximately one billion years to have a 0.1% chance of repeating the same UUID — see e.g. Rehak (2017) for an informal illustration.

[5] https://rsync.samba.org/

[6] https://pydantic.dev/

2.3. Replicability, reproducibility, and extensibility

The importance of replicability and reproducibility of research has recently received significant attention across empirical fields, as it forms an essential part of the scientific method impacting the scientific community's ability to assess the validity and reliability of empirically backed claims (see e.g. Moonesinghe et al. 2007; Simons 2014). This also extends to the larger field of linguistics (see e.g. Sönnig & Werner 2021 and articles therein; Kobrock & Roettger 2023) both regarding the replicability of methods and the reproducibility of results given extant data that may not be replicable directly (see also Berez-Kroeker et al. 2019).

Research on language attitudes has to-date relied on non-standardised methods, where questionnaires and other methodologies are frequently modified and adapted on an ad-hoc basis without sufficient documentation and without materials and data being made fully available, impeding not only replicability and reproducibility but also the comparability of results across studies.

For instance, consider the following sample of five studies employing the Matched Guise Test (MGT): Lambert et al. (1960), Echeverria (2005), Soukup (2013), Loureiro-Rodriguez, Boggess & Goldsmith (2013), and Price & Tamburelli (2020). Of these, two employed bipolar oppositional adjective pairs, once along an oppositional scale (Soukup 2013) and once as singleton agree-disagree items (Price & Tamburelli 2020), while all others had only singular adjectives along an agree-disagree scale. Lambert et al. (1960) used a 6-point Likert-type scale, while all others used a 5-point scale. As shown in Appendix A, the studies presented 10, 12, 6, and 8 guises respectively, with varied numbers of speakers ranging from 2 to 4 (apart from fillers, on which reporting was generally lacking). In total, the five studies used 109 different traits, with individual studies asking participants to rate between 15 and 30 traits. Of the 109 traits across studies, none appeared in all five studies, and only confidence, intelligence, progressiveness, sense of humour, and trustworthiness appeared in four of the studies. 44 of the traits appeared only in a single study. This situation prevents us from meaningfully comparing results across studies and thus from assessing their cross-cultural/cross-linguistic validity. A hypothetical mini-meta-analysis of the five studies could at best compare only 5 traits, with 1-2 of them treated as 'NAs' for all except Soukup (2013). The practical comparability of a large body of work employing the MGT thus currently approaches zero. We propose that this can and must be addressed by a more standardised, freely available, and reusable implementation such as ours, which maximises constants across studies and populations.

Regarding reproducibility and replicability (the abilities to derive the same results from the same inputs and the ability to re-run a study [possibly with a new population sample] to arrive at comparable results, respectively) both present major challenges to the subfield at present. None of the five studies in the sample above shared either their data or means of analysis to a degree that would be sufficient to reproduce their results. None of the studies shared their protocol or materials (or described them in sufficient detail) and only one study (Price & Tamburelli 2020) gave an example of one of the rating sheets presented to participants, making reasonably faithful replications effectively unfeasible. While one would hope that most researchers would be willing to share data or more detail on materials and

protocol upon request (ignoring here issues such as researchers becoming inactive or not holding on to old files and materials), even then these might come in widely different formats, employ different measures and scales, rely on closed platforms or formats or otherwise unavailable technology, and so on. Our app addresses the data issue by adopting an easily shareable, validatable, consistent, open data format (based on JSON), with a one-file-per-response approach. Releasing the app as free and open-source software ("FOSS"), with any changes being version controlled and transparent, goes a significant way toward addressing the concerns regarding protocol, materials, and technological availability. If a researcher makes an adaptation or improvement, they can commit this to the public repository and the only information they need to share to make their study replicable is the version number and protocol that was followed with participants on the ground. To share data, they can simply upload the set of data files produced by the app to a public archive/repository such as OSF, and other researchers who have previously worked with output from the app can easily consume that data in the same manner.

In general, we believe that the less a researcher must do themselves to make their work replicable, reproducible, and comparable, the more likely it is that they will do so. This is the major methodological reason for not only developing but also adopting tools such as the L'ART Research Assistant.

Finally, let us say a word about extensibility. A researcher working on the attitudes of French—English bilinguals in Canada needs an MGT with (Canadian) French and English guises, but the app may not yet have these available. To increase the adoption of the app and the benefits this would bring (as discussed above), the implemented research tasks must be easily adaptable and extensible without compromising their consistency, openness and transparency. This is achieved in three ways. First, for tasks where all materials are text-based, such as the LSBQe or the AToL, all a researcher needs to do when intending to apply these in a new location for which suitable materials are not already available is to edit a single text-based localisation file. Second, for tasks that rely on additional stimuli (such as the audio stimuli of the MGT/AGT), researchers can simply include these stimuli along with their localisation file. Third, where a new task is to be implemented, or a wholesale change proposed for a task,[7] researchers can submit code based on the app's APIs to be integrated into the app — this is facilitated by full documentation and the use of a technology stack widely employed across the scientific community (Python, JavaScript, and HTML). In all cases, researchers can and should propose changes and/or additions to be included in the official repository on GitHub via pull requests or issues. This makes the process of incorporating changes and additions transparent as well as making them available to the wider research community. Instructions on how researchers can go about adding new translations and adaptations are available as part of the official documentation of the app, available at https://lart.readthedocs.io/projects/research-assistant/ (see also Breit et al. 2023 for

[7] Changes to existing tasks will only be accepted where they maintain compatibility with the wider use-case of that task — alternatively they can be added as new standalone tasks.

a more practice-oriented, less formal description of the app and the research tools implemented therein).

3. Research tasks currently available in the app

In this section, we describe the principal user-facing features of the L'ART Research Assistant app, namely the research tools, or tasks, that are currently available as part of the app, focusing here principally on technical and methodological aspects.

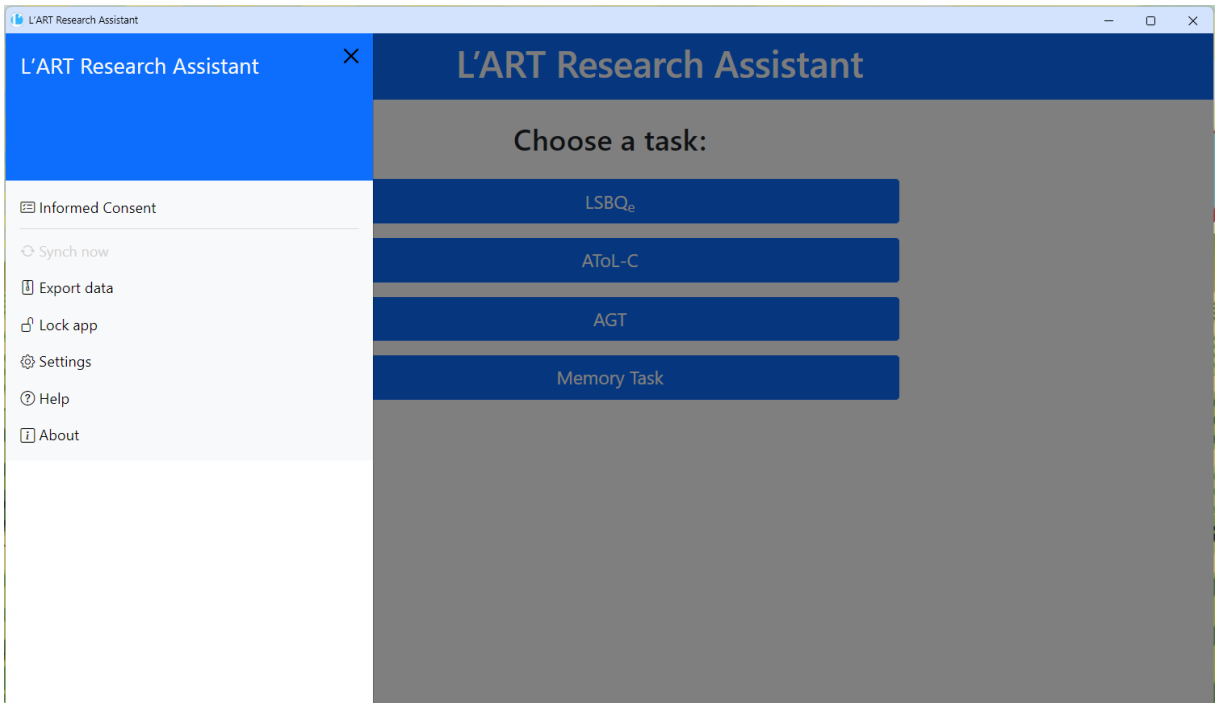

Figure 1: *App home screen with opened menu*

3.1. Informed consent

The built-in informed consent facility presents an optional entry point for data collection via the app. It presents users with a simple participant information sheet that shows list-based information (such as study title, purpose of the research, ethics approval), at the bottom of which participants find a checkbox to confirm that they have read and understood the information and that they agree to take part voluntarily. This section is followed by a short block with eligibility criteria, which likewise asks participants to confirm that they are eligible to take part in the study, as illustrated in Figure 2.

Researchers can fully customise the information displayed, both for the informed consent section and the eligibility criteria, by editing a simple text-based file containing a list of the items to display to participants to include specific descriptions and other relevant information for their project.

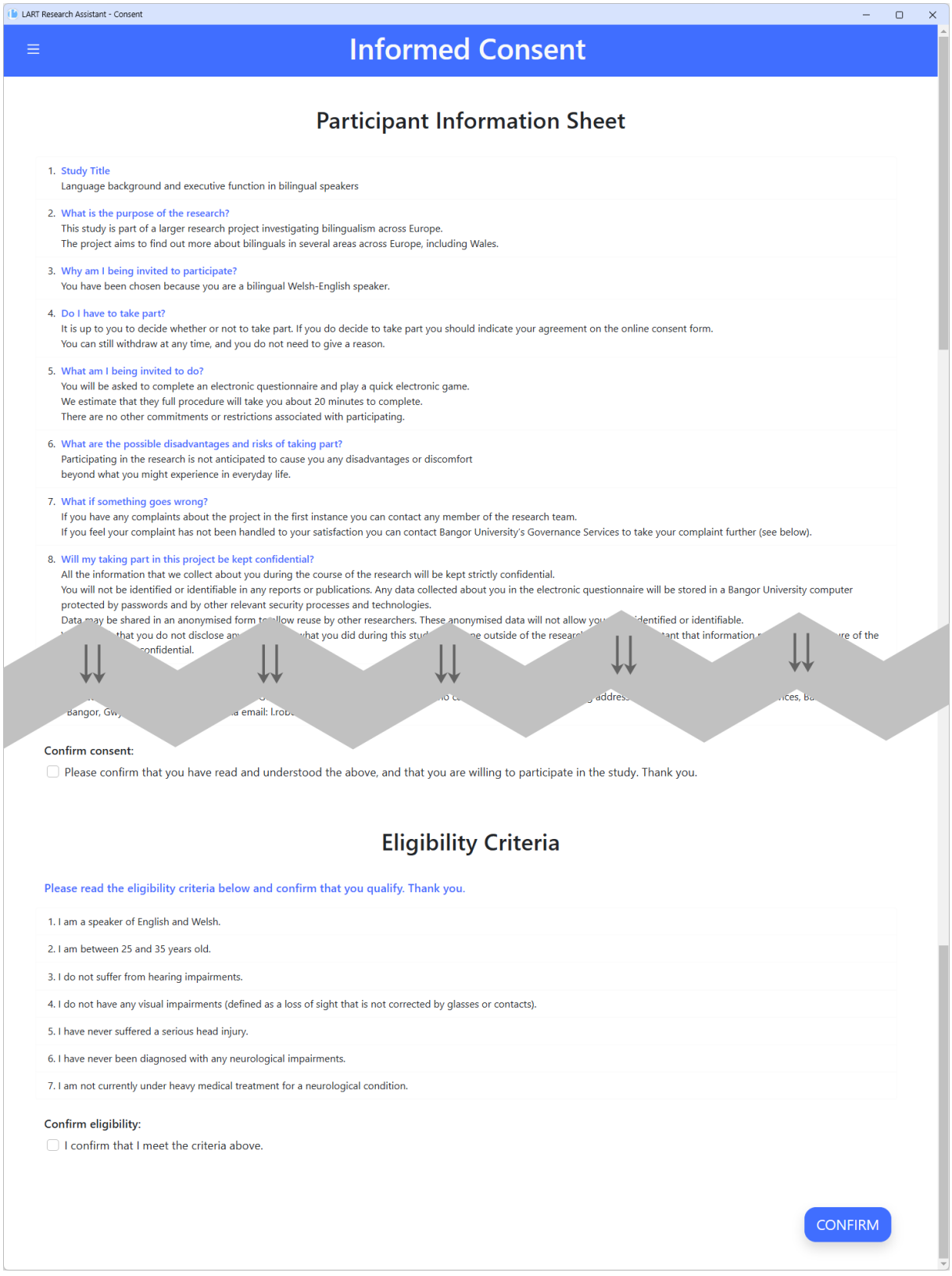

Figure 2: *Example of a digital informed consent form. Shown are the top portion (above the grey divider) with customisable project information and the bottom portion of the informed consent screen (below the grey divider) with customisable eligibility criteria confirmation.*

Importantly, apart from the two checkboxes, the only information solicited by the informed consent facility is the participant ID which the researcher pre-assigns to the participant. This means that no sensitive personal information (e.g. name, email address, signature) is recorded, so that data files with the participants' consent record can be shared along with other data, without special precautions needed for data protection. While we are aware that this will be incompatible with some use cases, principally where local ethics committees mandate that the identity of a participant is recorded on the consent form, we did not want to compromise the principle of avoiding the collection of data that would regularly need to be sanitised before being shared or stored on unprotected hardware to conform to common data protection regulations.[8] In those cases, researchers can of course simply continue to employ regular paper-based consent forms without impacting the core functionality offered by the app.

3.2. LSBQe: Language and Social Background Questionnaire

The LSBQe ('e' for electronic) is an adapted version of Anderson et al.'s (2016, 2018) Language and Social Background Questionnaire (LSBQ). Adaptations were made with two goals in mind: first, to increase the domain of applicability of the questionnaire by making it less task- and location-specific; second, to take advantage of the additional tools on offer in a digital implementation compared to a paper-based presentation.

Adaptations of the first kind include removal of "the other language" where that language could be predetermined, for example based on the specific localisation or from prior user input; the removal of some eligibility-based questions that were specifically designed with particular research paradigms in mind (e.g. neurophysiological studies); a neutral phrasing for questions of origin and past places of residence (where the original assumes Canada as a base and asserts this in its phrasing); and a re-design and re-referencing of educational background questions to 5 levels referenced to the European Qualifications Framework (EQF), as follows: (1) EQF Level 1, (2) EQF Levels 2-3, (3) EQF Level 4, (4) EQF Levels 5-6, and (5) EQF Levels 7-8. Different localisations include appropriate examples for each level based on the location of the research, such as "Abschluss der Grund– oder Primarschule, CEB, getuigschrift basisonderwijs, oder weniger" for EQF Level 1 in the version localised for German in Belgium. The EQF was chosen because it is easily cross-referenceable across different countries and education systems, can be compared with the International Standard Classification of Education (ICSD), and is recognised outside Europe as a vital reference framework for comparing qualifications worldwide (Chakroun 2010).

[8] Particularly the principle of data minimisation in Art. 5(1)(c) of the General Data Protection Regulation.

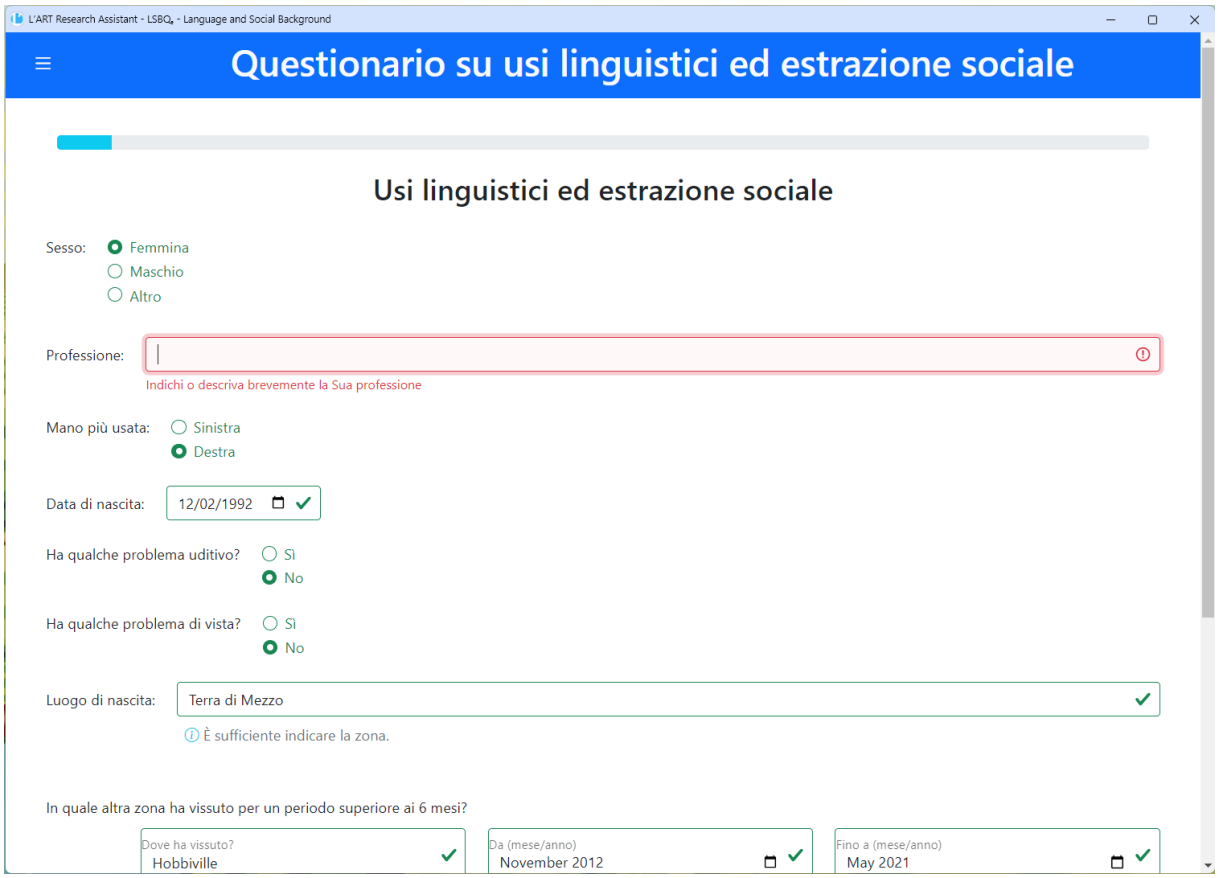

Figure 3: *First page of the LSBQe, showing active input validation (Italian)*

Technological adaptations and improvements include page-wise input validation with feedback (as shown in Figure 3), the automated display/hiding of relevant/irrelevant questions based on previous answers (e.g. in Figure 3, if the participants selects 'Yes' to the question on visual impairment, an additional question appears beneath asking whether they use glasses or contact lenses, but remains hidden otherwise), the expandability of repeatable groups of questions such as the list of places a participant has lived in (so there is no chance of a participant running out of a pre-determined maximum length of lists, and they can also not miss partial questions relevant to each), the requirement to give at least two languages for bilingual participants, the suggestion of possible locally relevant languages when entering languages they speak (as illustrated in Figure 4), and the use of continuous sliders (Figure 5) instead of Likert-type scales where continuous range data is to be collected (e.g. questions asking how proficient a participant is on proficiency in speaking and understanding a particular language, ranging from 'No Proficiency' to 'High Proficiency'), and the requirement to mark non-applicable questions explicitly as not applicable for validation purposes, which ensures questions are not accidentally left blank. The use of sliders allows to measure a response as a double-precision floating point number between 0 and 100, enabling much finer-grained data without burdening the participant with excessive options. Additionally, the resultant data is to some extent insulated from the "scale coarseness effect" (Symonds 1924; Russell & Bobko 1992; Aguinis, Pierce & Culpepper 2009) – a methodological artifact which causes a downward bias in correlation coefficients, arising from measurement of a continuous variable via a categorical scale, such as Likert-type scales.

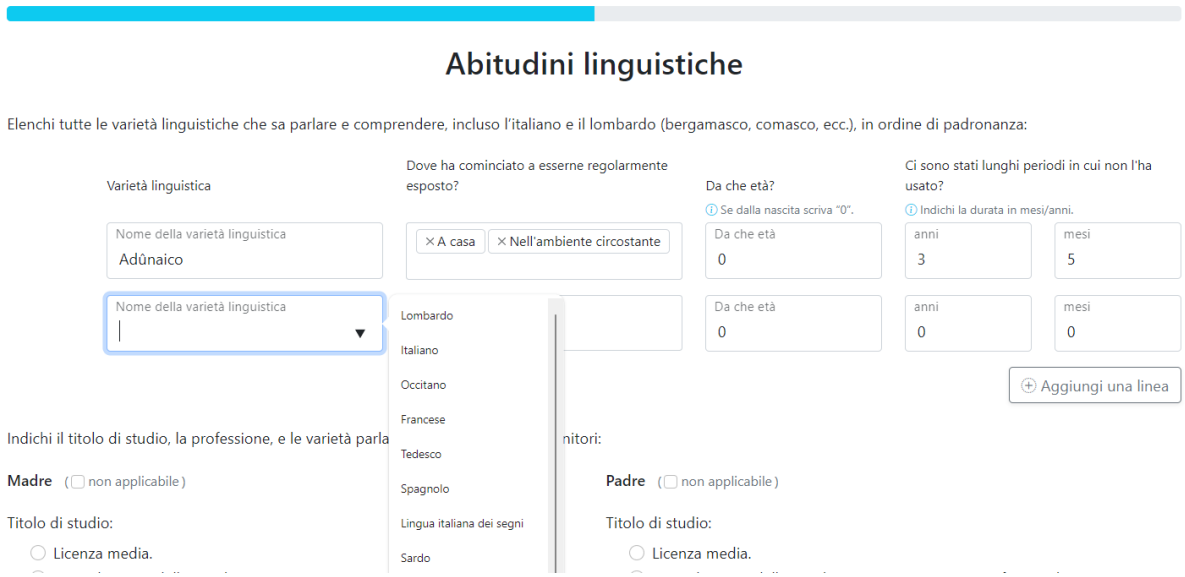

Figure 4: *Language entry fields on the LSBQe with suggested languages as the participant types and the option to add additional language entry fields (Italian)*

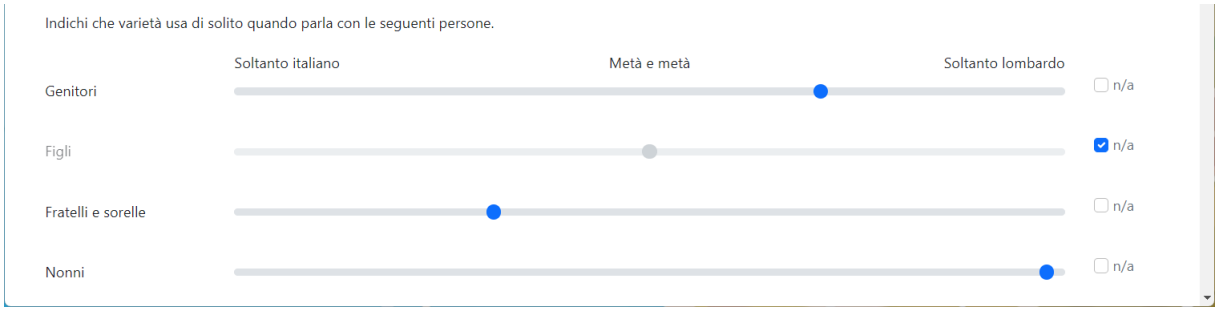

Figure 5: *Sliders and explicit "not applicable" options for optional, continuous range data collection (Italian)*

Our final modification was to include a final open ended free-text question where participants can optionally enter any additional information they believe might be relevant. In all other respects, we have attempted to make the LSBQe as faithful to the design of the established LSBQ as possible.

3.3. AToL-C: Attitudes Towards Languages

The AToL-C ('C' for continuous) is our implementation of Schoel et al.'s (2013) Attitudes Towards Languages (AToL) Scale, a cross-linguistically well-validated attitudinal questionnaire employing 15 oppositional adjective pairs yielding three primary measurement dimensions, namely: Value, Structure, and Sound. Structure and Sound capture scale items principally describing the structural and sonic aspects of the language, respectively. Value is a hierarchical superordinate factor, which although shaped by its own subscale, shows correlation to both Sound and Structure (see Schoel et al.: 24–27).

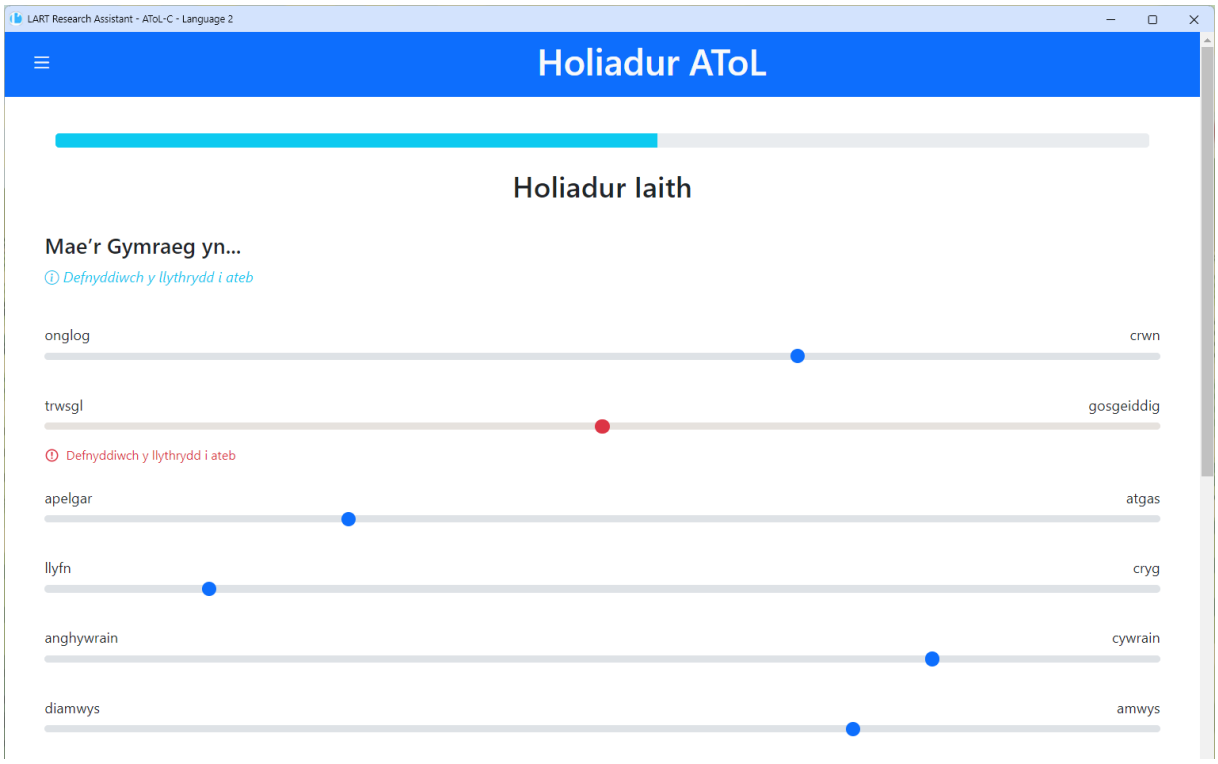

Figure 6: *AToL-C with continuous sliders and user feedback for a slider that was not moved by the participant (Welsh)*

Because Schoel et al. (2013) went through an extensive selection and validation process of these adjectives and dimensions, our implementation seeks to be as faithful as possible to the original. However, as shown in Figure 6, we take advantage of the digital format by adopting continuous sliders instead of the 5-point Likert-type scales used in the original (see 3.2 for more on continuous sliders), and by introducing randomisation of the adjective pairs' presentation order across participants. Neither of these should negatively affect the validity or reliability of the AToL. For the same reason, and because we see the establishment of well-validated reference tools as especially important in our field, we also chose not to implement extensions or variations of the AToL, such as Lehnert et al.'s (2016, 2018) extended AToL or Attitudes Towards Language Users Scale (AToLU), though they can easily be derived and implemented from our AToL codebase.

3.4. AGT: Verbal/Matched/Audio Guise Task

The speaker evaluation paradigm involves exposing participants to different audio-recorded guises representing different linguistic varieties that participants must rate (Dragojevic & Goatley-Soan 2022). The design we have adopted (which we generically term the Audio Guise Task; AGT) allows researchers to implement the speaker evaluation paradigm both using the Matched Guise Test (MGT; Lambert et al. 1960) and the Verbal Guise Technique (VGT;

Markel, Eisler & Reese 1967).[9] Three fully functional MGT localisations are currently included in the app, for the language pairs Welsh–English, Moselle-Franconian–German, and Lombard–Italian.

Following standard procedure for both the MGT and VGT, the AGT asks participants to complete an evaluative questionnaire rating speakers on a range of traits. This includes 18 traits (see Appendix B), which were selected based on a cross-comparison of several published MGT studies (see also the discussion in Section 2.3). In constructing the trait list, we also considered whether they (intuitively) aimed at status or solidarity,[10] the valence of a trait,[11] and translatability across the languages we implemented, with at least one member of our team possessing linguistic expertise in each of these languages.

Figure 7: *Guise presentation and trait rating during the AGT (German)*

[9] The chief difference between the MGT and the VGT are that the former has two guises each (one per language) from six speakers, whereas the verbal guise technique uses twelve speakers providing one sample each (across two or more languages).

[10] For data analysis, this must be decided on a per-population basis, e.g. via principal factor analysis.

[11] We included items which have the opposite valence direction (*agree* is worse) to the majority (*agree* is better), while keeping the scale directionality constant (disagree→agree) to guard from potential scale direction effects (Yan, Keusch & He 2018; Salzberger & Koller 2019). This principle can also be observed in Schoel et al.'s (2013) bipolar AToL traits.

Procedurally, participants are first presented with instructions followed by a practice guise, to familiarise them with the task and the traits they will be rating. This is important for the AGT given that participants only have a limited time to hear the test guises. They are then presented successively with 12 guises (including 4 filler guises) in pseudo-randomised order.[12] The trait list presented is randomised afresh for each guise, and while participants do not have an upper time limit, they must complete all 18 ratings and listen to at least half the guise (or 30 seconds of the guise, whichever is shorter) before they can advance. This minimum time enforcement is to prevent participants from rating guises too quickly, e.g. before they have had a reasonable opportunity to develop an opinion toward the guise. Participants cannot relisten to a guise or return to earlier guises, but they are provided with visual feedback on the remaining playtime of the current guise (see Figure 7).

3.5. Memory Task

We also implemented a short memory game, adapted from Tarnate (2022). This task is intended as a general distractor that can be employed when running tasks in series. This is particularly relevant in research on language attitudes, where it is standard practice not to fully disclose the aim of the study to participants in order to render attitudinal measures less direct (see e.g. Pharao & Kristiansen 2019). This serves to minimise acquiescence bias, where participants tend to give the response they believe the researchers are looking for (Jackson & Messick 1965), and the social desirability effect, where participants respond with the attitude they think is perceived more desirable (Diekmann 2007).

[12] The randomisation is constrained such that fillers are spaced regularly, that repeating speakers are maximally spaced apart, and that the same language does not repeat in more than two successive test guises (see Breit et al. 2023: 20–21 for more detail).

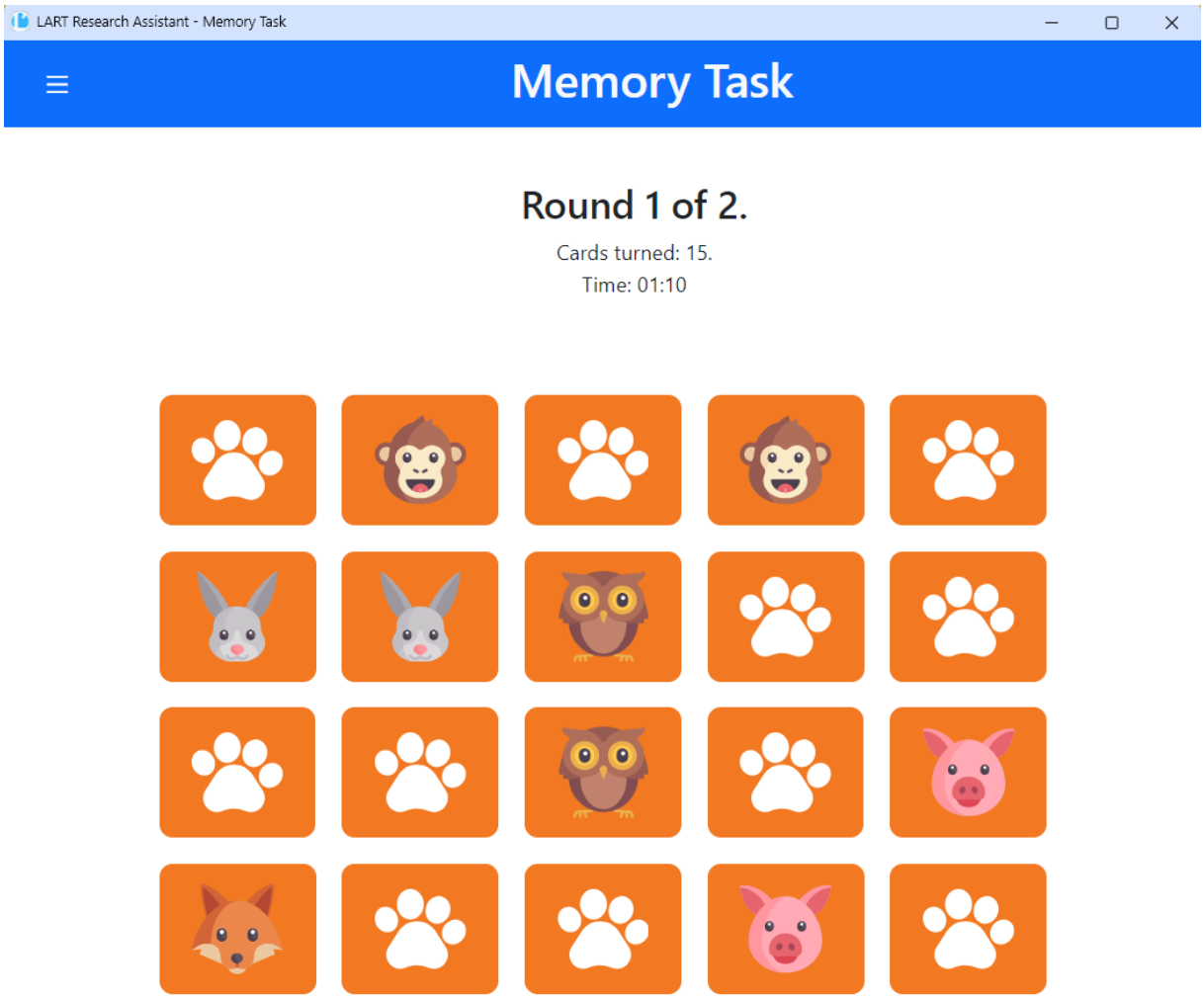

Figure 8: *The Memory Task in the L'ART Research Assistant, showing the cards turned counter, timer, and some uncovered/matched card pairs (illustrations on playing cards © Freepik at flaticon.com, licensed under CC BY 3.0)*

The game (shown in Figure 8) requires participants to successively uncover and match identical card pairs in the shortest amount of time. Participants are scored on time and number of cards uncovered. Participants play the game twice and are instructed on the second attempt to try and beat their previous score.

4. Case study

In this section, we report preliminary results from a larger project currently underway, which investigates attitudes across three bilingual language communities in three countries, where each of the regional/minority languages is found in a different sociopolitical and sociolinguistic context:

- Welsh–English in North Wales: official recognition and active institutional support, concerted effort to increase speaker numbers;

- Moselle-Franconian–German in East Belgium: not officially recognised, indirectly supported via Standard German, diglossia with slow language shift toward German; and
- Lombard–Italian in Northern Italy: not officially recognised, a case of "benign neglect" (Coluzzi, 2009; Coluzzi et al., 2018; see also "no-policy policy" in Fishman 2006: 325), advanced language shift toward Italian.

Our primary focus here is on demonstrating the use of the L'ART Research Assistant in fieldwork, which is why we omit much of the otherwise warranted background and discussion around the rationale, linguistic situation and implications of these studies.

4.1. Methods

4.1.1. Participants

Our population sample included early bilinguals aged 24-36, the 'current parent generation' most influential on intergenerational transmission. The sample sizes for the AToL were 42 (female = 21, mean age = 28.0 in Wales, 30 (female = 22, mean age = 29.63) in Belgium, and 39 (female = 24, mean age = 28.62) in Italy. For the MGT these were 44 (female = 19, mean age = 28.20) in Wales, 40 (female = 30, mean age = 29.56) in Belgium, and 40 (female = 23, mean age = 30.18) in Italy.

4.1.2. Materials

We used the L'ART Research Assistant (Version 0.5.0; see Section 3), with the LSBQe as background questionnaire, combined with either the AToL-C and the Memory Task, or the AGT. All tasks were run as implemented for the respective language pairs in that version of the app, which includes all our materials.

4.1.3. Procedure

For the AToL, participants were told they would take part in a study on executive function in bilinguals comparing linguistic communities across Europe. Participants' informed consent and eligibility was recorded with the app's Informed Consent facility (Section 3.1). They were told we would first collect comprehensive information on their linguistic background, before they would complete a short memory task to assess their executive function. They then completed the LSBQe (Section 3.2), followed by the AToL-C (Section 3.3) and finally the Memory Task (Section 3.5) as a distractor. This took about 30 minutes total.

For the MGT, participants were told they would take part in a study looking at how people from different bilingual backgrounds across Europe perceive and rate different voices. Informed consent, eligibility, and background questionnaire were as above, with the LSBQe being followed by the AGT (Section 3.4). No distractor task was employed this time. This took about 30 minutes total.

For both studies, experiments were administered in pre-booked rooms across various locations. Due to the practicalities of our locations, the researcher was present either in the room or just outside, but participants were situated such that they were focused on the computer where they completed the task and could not observe the researcher. They were told the researcher would be completing paperwork in the meantime but would be available to assist them if necessary. Participants were compensated for their time at the end of their participation.

Data analysis was carried out in IBM SPSS (Version 29). The AToL was analysed using a 3×2 (community: Wales, Belgium, Italy × language: majority, minority) factorial analysis of variance (ANOVA) with repeated measures on the third factor. For the MGT, we first conducted factor analysis using principal axis factor (PAF) with varimax (orthogonal) rotation of the 18 traits (see Appendix B) across all participants, with the Kaiser-Meyer-Olkin measure of sampling adequacy examined to confirm factorability, successively excluding outliers or single-trait factors until we arrived at two factors consistent with the status and solidarity dimensions, employing Cronbach's alpha to confirm reliability of the final factorisation. The MGT ratings were then analysed using a 2×2 (dimension: solidarity, status × language: majority, minority) factorial ANOVA with repeated measures and with community (Wales, Belgium, Italy) as a 3-level between-subjects factor.

The data presented here are freely available from the Open Science Framework (https://doi.org/10.17605/OSF.IO/CP6RE); see also Brasca et al. (2024).

4.2. Results: AToL

Statistical analysis revealed significant main effects for community ($F(4,214) = 27.904$, $p < .001$, $\eta_p^2 = .343$) and for AToL dimension (Value, Sound, Structure) ($F(5,105) = 28.93$, $p < .001$, $\eta_p^2 = .524$). We also found an AToL dimension × community interaction ($F(8,210) = 12.228$, $p < .001$, $\eta_p^2 = .318$), indicating that the mean score for each AToL dimension was different across communities. Within-subject tests showed that the difference across AToL dimensions was significant for both majority ($F(1,108) = 9.685$, $p = .002$, $\eta_p^2 = .082$) and minority languages ($F(1,108) = 13.547$, $p < .001$, $\eta_p^2 = .111$). Between-subject effects were also significant for both majority ($F(2,108) = 40.071$, $p < .001$, $\eta_p^2 = .426$) and minority languages ($F(2,108) = 16.212$, $p < .001$, $\eta_p^2 = .231$).

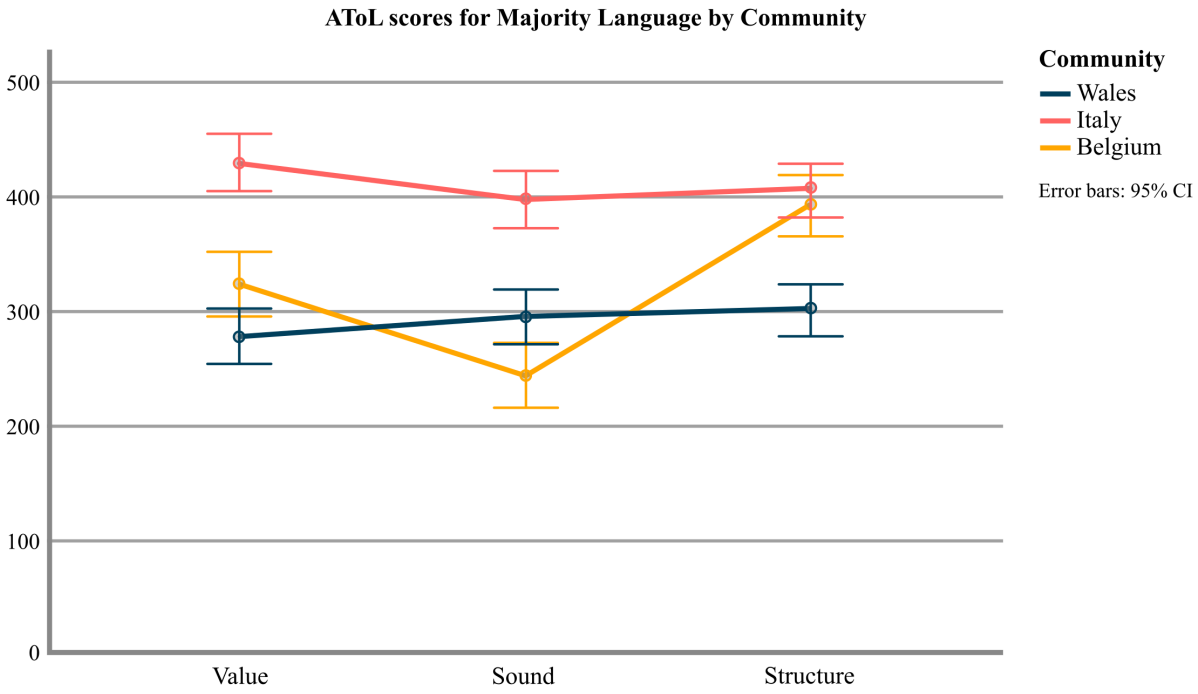

Figure 9: *AToL scores for Majority Language (English, Italian, German) by Community*

Figure 9 shows the AToL scores in each dimension for the respective majority languages (English, Italian, German). As can be seen there, the Lombard–Italian speakers rated Italian extremely high across dimensions, while the Welsh–English and Moselle-Franconian–German rated English and German medium high for Value and Sound, though notably for German, Structure was rated much higher.

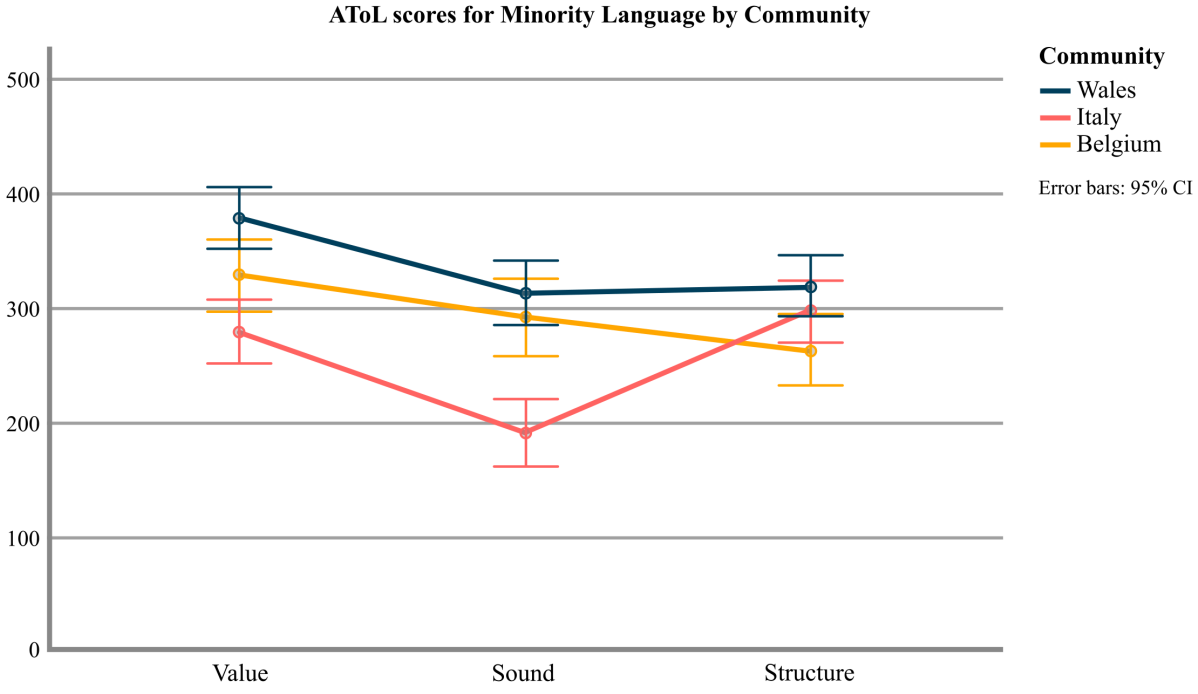

Figure 10: *AToL scores for Minority Language (Welsh, Lombard, Moselle-Franconian) by Community*

A different pattern is revealed by Figure 10, showing the analogous scores for the minority languages (Welsh, Lombard, Moselle-Franconian). Most notably, the Lombard–Italian speakers rated Lombard low to medium high, and particularly poorly in the Sound dimension. This stands in stark contrast to their rating of Italian. The Welsh–English speakers rated Welsh much higher than English in the Value dimension, but similarly to English in Sound and Structure. Compared to German, the Moselle-Franconian–German speakers rated Moselle-Franconian similar in Value, slightly better in Sound, and worse in Structure.

4.3. Results: MGT

Statistical analysis revealed significant main effects for language ($F(1,121) = 19.074$, $p < .001$, $\eta_p^2 = .136$) and MGT component (solidarity, status) ($F(1,121) = 367.723$, $p < .001$, $\eta_p^2 = .752$). We also found interactions of MGT component × language ($F(1,121) = 85.931$, $p < .001$, $\eta_p^2 = .415$) and community × language ($F(1,121) = 21.575$, $p < .001$, $\eta_p^2 = .263$), indicating that mean MGT scores were different for languages by community and by MGT component. Within-subject tests showed that the difference between MGT components was significant for language (majority, minority) ($F(1,121) = 19.074$, $p < .001$, $\eta_p^2 = .136$). There was a significant between-subject effect for community ($F(2,121) = 14.17$, $p < .001$, $\eta_p^2 = .190$).

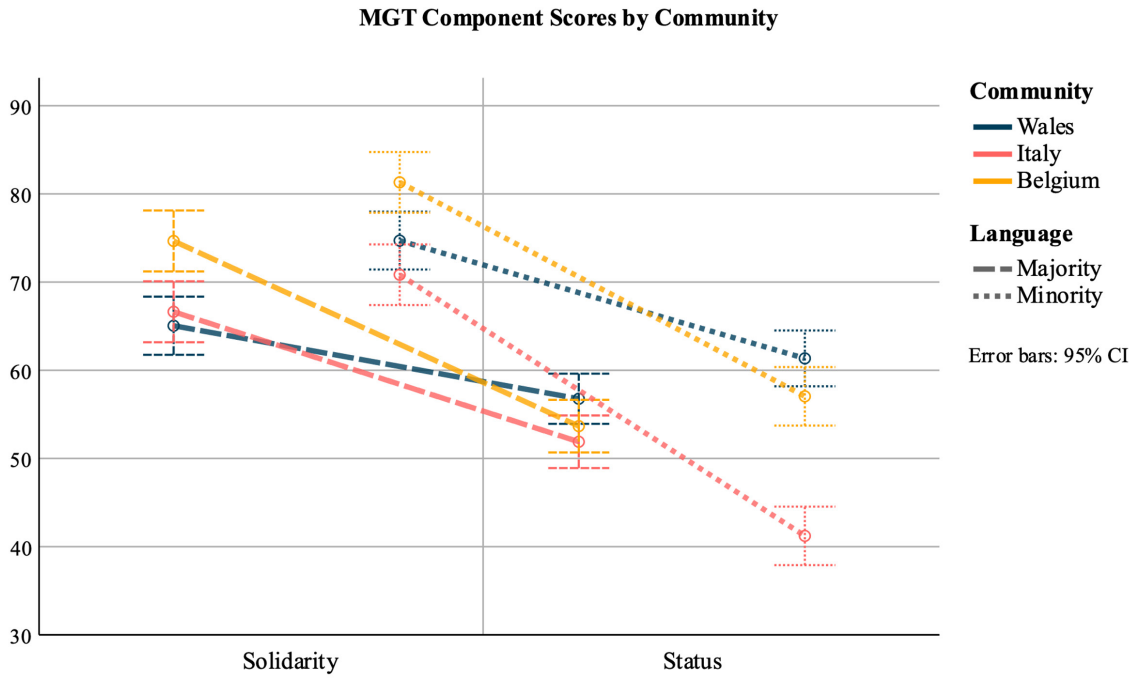

Figure 11: *MGT scores by component (solidarity, status) for majority and minority languages across the three communities*

As can be seen from Figure 11, showing mean MGT scores by MGT component and language for each of the three communities, the minority languages score consistently higher in solidarity compared to the majority language, showing that at least Welsh and Moselle-Franconian engender greater solidarity — however, this is much less pronounced for Italian v. Lombard, with a notable overlap in their 95% CI. For the status component, we see that

Lombard scores much worse than Italian, indicating a significant detriment in the perceived status. Moselle-Franconian scores slightly better than German on status, though there is a notable overlap in the 95% CI here. Welsh also scores higher than English on the status dimension, with a smaller overlap of 95% CI compared to Moselle-Franconian–German, possibly indicating that Welsh may outstrip English in both solidarity and status in our sample.

4.4. Discussion

The results from our AToL and MGT studies show what might be expected based on the vitality of the respective minority languages and the socio-political setting of the three communities: Attitudes towards Lombard are consistently the lowest, it performs particularly low on the status component of the MGT and in the AToL Sound dimension, both compared to Italian and the other minority languages. This reflects the fact that it is the most challenged of these languages both socio-politically and in terms of its declining vitality. Conversely, Welsh has seen significant political support in Wales since devolution, and performs better than English across all the AToL dimensions and MGT components, with a particular boost in the AToL's Value dimension. Moselle-Franconian exemplifies what one might expect in a fairly stable diglossic situation engendering functional separation, where both languages perform similarly overall but are differentiated mainly in the individual components, where Moselle-Franconian performs particularly well in the MGT's solidarity component and particularly poorly, compared to German, in the AToL's Structure dimension.

Beyond this, the comparison of results from the AToL and MGT highlights another aspect of interest regarding these measures. Namely, while overall results by community are fairly consistent, the factors which vary most significantly for a given language are not necessarily commensurate. For example, German has its highest score in the AToL Structure dimension, which is not reflected or transferable to the MGT, while Moselle-Franconian is associated with a significant solidarity boost in the MGT, which is not apparent from any of the AToL dimension scores. Further research needs to address to what extent these are measures of the same attitudinal object, and to what extent attitudinal research might need to move not only toward adopting an approach employing a multimethodological battery of tests, but also address the contribution and significance of the different kinds of measures toward an overall attitudinal assessment.

5. Conclusion

In this paper we introduced a new digital app, the L'ART Research Assistant, which aims to assist researchers in conducting attitudinal and other studies on multilingual populations, especially where this involves a majority and a regional/minority/minoritised/heritage language. We discussed several aspects of research methodology in the field that can benefit from the opportunities offered by the adoption of such technology. The principal strengths highlighted were that adoption of this tool can improve the transparency and openness of

materials and procedures, the replicability and reproducibility of studies, and the consistency and comparability of results within and across different studies.

The tasks currently implemented in the app reflect the most common methodological needs in research on language attitudes, covering the collection of informed consent, comprehensive participant background information, questionnaire-based direct measures of attitudes via the AToL, speaker-evaluation paradigm measures of attitudes via the AGT (MGT/VGT), and a basic distractor task.

We demonstrated the use of the app and its tools through a case study of attitudes toward a majority and a minority language conducted across three bilingual populations, which showed how the benefits discussed materialise in practice: methods and protocols are transparent and shared easily and the results are readily comparable in analysis, enabling us to draw inferences both about differences across the populations studied and the methods employed. Wider adoption of the app, and contributions in the forms of further localisations and new tasks, could bring these benefits to bear on the wider research community in multilingualism, language attitudes, and language maintenance.

Acknowledgments

We would like to thank colleagues at Bangor and beyond who have provided feedback and assistance during the development of the L'ART Research Assistant, particularly Jago Williams for the initial testing of the app and Chloe Cheung for her contribution to the app's online documentation, as well as the anonymous reviewers for their helpful comments, and the audiences at various conferences who furnished us with insightful discussions as the work presented here was in progress. We are grateful to the many local helpers in Wales, Belgium, and Lombardy who assisted with facilitation of the research discussed in Section 4. This work was supported by the Economic and Social Research Council [grant number ES/V016377/1].

References

Aguinis, H., C. Pierce, and S. Culpepper. 2009. Scale coarseness as a methodological artifact: Correcting correlation coefficients attenuated from using coarse scales. *Organizational Research Methods* 12(4): 623–652. https://doi.org/10.1177/1094428108318065

Anderson, J. A., L. Mak, A. Keyvani Chahi, and E. Bialystok. 2016. The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. Supplementary materials (Version 1). *figshare*. https://doi.org/10.6084/m9.figshare.3972486.v1

Anderson, J. A., L. Mak, A. Keyvani Chahi, and E. Bialystok. 2018. The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior Research Methods* 50(1): 250–263. https://doi.org/10.3758/s13428-017-0867-9

Berez-Kroeker, A. L., L. Gawne, S. Smythe Kung, B. F. Kelly, T. Heston, G. Holton, P. Pulsifer, D. I. Beaver, S. Chelliah, S. Dubinsky, R. P. Meier, N. Thiberger, K. Rice, and A. C. Woodbury. 2019. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1): 1–18. https://doi.org/10.1515/ling-2017-0032

Brasca, L., M. Tamburelli, I. Gruffydd, and F. Breit. 2024. A matter of strength: Language policy, attitudes, and linguistic dominance in three bilingual communities. *Journal of Multilingual and Multicultural Development*. https://doi.org/10.1080/01434632.2024.2408448

Breit, F., M. Tamburelli, I. Gruffydd, and L. Brasca. 2023. *The L'ART Research Assistant: A digital toolkit for bilingualism and language attitude research* [Manuscript]. Bangor University. https://doi.org/10.13140/RG.2.2.18801.35686

Chakroun, B. 2010. National qualification frameworks: From policy borrowing to policy learning. *European Journal of Education* 45(2): 199–216. https://doi.org/10.1111/j.1465-3435.2010.01425.x

Coluzzi, P. 2009. Endangered minority and regional languages ('dialects') in Italy. *Modern Italy* 14(1): 39–54. https://doi.org/10.1080/13532940802278546

Coluzzi, P., L. Brasca, M. Trizzino, and S. Scuri. 2018. Language planning for Italian regional languages: The case of Lombard and Sicilian. In D. Stern, M. Nomachi, and B. Belić (eds.), *Linguistic Regionalism in Eastern Europe and Beyond: Minority, Regional and Literary Microlanguages,* 274–298. Peter Lang. https://doi.org/10.3726/b14464

Diekmann, A. 2007. *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen*. Rowohlt.

Dragojevic, M., and S. Goatley-Soan. 2022. Americans' attitudes towards foreign accents: evaluative hierarchies and underlying processes. *Journal of Multilingual and Multicultural Development* 43(2): 167–181. https://doi.org/10.1080/01434632.2020.1735402

Echeverria, B. 2005. Language Attitudes in San Sebastian: The Basque Vernacular as Challenge to Spanish Language Hegemony. *Journal of Multilingual and Multicultural Development* 26: 249–264. https://doi.org/10.1080/01434630508668407

Fishman, J. A. 2006. Language policy and language shift. In T. Ricento (ed.), *An Introduction to Language Policy: Theory and Method*, 311–328. Blackwell.

Grieve, J. 2021. Observation, experimentation, and replication in linguistics. *Linguistics* 59(5): 1342–1356. https://doi.org/10.1515/ling-2021-0094

Jackson, P. W., and S. Messick. 1965. The person, the product, and the response: Conceptual problems in the assessment of creativity. *Journal of Personality* 33(3): 309–329. https://doi.org/10.1111/j.1467-6494.1965.tb01389.x

Kircher, R., and L. Zipp (eds.) 2022. *Research methods in language attitudes*. Cambridge University Press. https://doi.org/10.1017/9781108867788

Kobrock, K., and T. B. Roettger. 2023. Assessing the replication landscape in experimental linguistics. *Glossa Psycholinguistics* 2(1): X. https://doi.org/10.5070/G6011135

Lambert, W. E., R. C. Hodgson, R. C. Gardner, and S. Fillenbaum. 1960. Evaluational reactions to spoken language. *Journal of Abnormal and Social Psychology* 60(1): 44–51. https://doi.org/10.1037/h0044430

Lehnert, T. E., S. Krolak-Schwerdt, and T. Hörstermann. 2016. Judging people and their language use: Evidence for the distinctness of attitudes towards languages and speakers' nationality. *Proceedings of EAPRIL 2015* 2: 262–274.

Lehnert, T. E., S. Krolak-Schwerdt, and T. Hörstermann. 2018. Language and nationality attitudes as distinct factors that influence speaker evaluations: Explicit versus implicit attitudes in Luxembourg. *Language & Communication* 61: 58–70. https://doi.org/10.1016/j.langcom.2018.01.005

Loureiro-Rodriguez, V., M. M. Boggess, and A. Goldsmith. 2013. Language attitudes in Galicia: Using the matched-guise test among high school students. *Journal of Multilingual and Multicultural Development* 34: 136–153. https://doi.org/10.1080/01434632.2012.729591

Markel, N. N., R. M. Eisler, and H. W. Reese. 1967. Judging personality from dialect. *Journal of Verbal Learning and Verbal Behavior* 6(1): 33–35. https://doi.org/10.1016/S0022-5371(67)80044-6

Moonesinghe, R., M. J. Khoury, and A. C. J. W. Janssens. (2007) Most Published Research Findings Are False—But a Little Replication Goes a Long Way. *PLoS Medicine* 4(2): e28. https://doi.org/10.1371/journal.pmed.0040028

Pantos, A. 2019. Implicitness, automaticity, and consciousness in language attitudes research. *Linguistics Vanguard* 5(s1). https://doi.org/10.1515/lingvan-2018-0007

Pharao, N., and T. Kristiansen. 2019. Reflections on the relation between direct/indirect methods and explicit/implicit attitudes. *Linguistics Vanguard* 5(s1). https://doi.org/10.1515/lingvan-2018-0010

Price, A., and M. Tamburelli. 2020. Welsh-language prestige in adolescents: Attitudes in the heartlands. *International Journal of Applied Linguistics* 30(2): 195–213. https://doi.org/10.1111/ijal.12274

Rehak, K. (2017, May 28). *Are UUIDs really unique?* Towards Data Science. https://towardsdatascience.com/are-uuids-really-unique-57eb80fc2a87

Russell, C. J., and P. Bobko. 1992. Moderated regression analysis and Likert scales: Too coarse for comfort. *Journal of Applied Psychology* 77(3): 336–342. https://doi.org/10.1037/0021-9010.77.3.336

Salzberger, T., and M. Koller. 2019. The direction of the response scale matters — accounting for the unit of measurement. *European Journal of Marketing* 53(5): 871–891. https://doi.org/10.1108/EJM-08-2017-0539

Schoel, C., J. Roessel, J. Eck, J. Janssen, B. Petrovic, A. Rothe, S. C. Rudert, and D. Stahlberg. 2013. 'Attitudes Towards Languages' (AToL) Scale: A Global Instrument. *Journal of Language and Social Psychology* 32(1): 21–45. https://doi.org/10.1177/0261927X12457922

Simons, D. J. 2014. The value of direct replication. *Perspectives on Psychological Science* 9(1): 76–80. https://doi.org/10.1177/1745691613514755

Sönnig, L., and V. Werner. (eds.) 2021. The replication crisis: Implications for linguistics [Special issue]. *Linguistics* 59(5).

Soukup, B. 2013. On matching speaker (dis)guises — revisiting a methodological tradition. In T. Kristiansen & S. Grondelaers (eds.), *Language (des)standardisation in late modern Europe: Experimental studies*, 267–285. Oslo: Novus.

Symonds, P. M. 1924. On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology* 7(6): 456–461. https://doi.org/10.1037/h0074469

Tarnate, J. 2022. *Memory Game* [Software]. CodePen. https://codepen.io/jeytii/pen/QoagLr

Vergnaud, A. C., M. Touvier, C. Méjean, E. Kesse-Guyot, C. Pollet, A. Malon, K. Castetbon, and S. Hercberg. 2011. Agreement between web-based and paper versions of a socio-demographic questionnaire in the NutriNet-Santé study. *International Journal of Public Health* 56(4): 407–417. https://doi.org/10.1007/s00038-011-0257-5

Yan, T., T. Keusch, and L. He. 2018. The impact of question and scale characteristics on scale direction effects. *Survey Practice* 11(2). https://doi.org/10.29115/SP-2018-0008

Appendix A: MGT traits used across 5 studies

Trait	Lambert et al. (1960)	Echeverria (2005)	Soukup (2013)	Loureiro-Rodriguez et al. (2013)	Price & Tamburelli (2020)	Usage Count
confidence	self-confidence	confident	self-confident		confident	4
intelligence	intelligence	intelligent	intelligent	intelligent		4
progressiveness		progressive	conservative	conservative	traditional	4
sense of humour	sense of humour		good sense of humour	to have a sense of humour	good-humoured	4
trustworthiness	(trustworthiness)	trustworthy	trustworthy	trustworthy		4
ambitiousness	ambition	ambitious		ambitious		3
attractiveness	good looks	attractive		attractive		3
education		cultivated	educated	educated		3
industriousness		hard-working	industrious	hard-working		3
kindness	kindness	kind		kind		3
likeability	general likeability		likeable		popular	3
openness		open		open-minded	easy-going	3
amusingness		amusing		amusing		2
boringness				boring	boring	2
competence			competent	apt		2
coolness		"cool"			cool	2
funness	entertainingness				fun	2
fashionability				modern	fashionable	2
friendliness			friendly		friendly	2
honesty			honest		genuine	2

leadership	leadership	leader				2
naturalness			natural		natural	2
politeness			polite	polite		2
sociability	sociability			introverted		2
aggression			aggressive			1
amenability					goody-two-shoes	1
annoyingness					annoying	1
arrogance			arrogant			1
awkwardness					awkward	1
caringness				caring		1
cleverness			clever			1
character	character					1
coarseness			coarse			1
conformity					conformist	1
cosmopolitanity				cosmopolitan		1
dependability	dependability					1
dullness					dull	1
efficiency				efficient		1
emotionality			emotional			1
excitingness					exciting	1
fakeness					fake	1
rebelliousness					rebellious	1
generosity		generous				1
goodness		good				1
groundedness					down-to-earth	1
tallness	height					1
humility		humble				1
ignorance				ignorant		1
impropriety				improper		1
jestfulness					likes a laugh	1
judgmentalness					judgmental	1
lameness					lame	1
nerdiness					nerdy	1
pretentiousness					pretentious	1
pride		proud				1
refinedness				refined		1
relatability					relatable	1
relaxedness			relaxed			1
religiousness	religiousness					1
roughness			rough			1
rusticity				rustic		1
seriousness			serious			1
smugness					smug	1
strictness			strict			1
studiousness					book-worm	1
unrefinedness				unrefined		1
uptightness					uptight	1
vulgarity				vulgar		1
Totals	**15**	**17**	**22**	**25**	**30**	**109**

Appendix B: AGT traits in the L'ART Research Assistant

Trait	Status / Solidarity (presumed)	Status / Solidarity (analysed)	Valence (higher score is …)
amusingness	solidarity	(excluded)	negative
pretentiousness	solidarity	solidarity	negative
friendliness	solidarity	solidarity	positive
honesty	solidarity	solidarity	positive
likeableness	solidarity	solidarity	positive
naturalness	solidarity	solidarity	positive
trustworthiness	solidarity	solidarity	positive
ignorance	status	solidarity	negative
ambitiousness	status	status	positive
attractiveness	status	status	positive
competence	status	status	positive
coolness	status	status	positive
educatedness	status	status	positive
influentiality	status	status	positive
intelligence	status	status	positive
internationality	status	(excluded)	positive
open-mindedness	status	status	positive
politeness	status	solidarity	positive
Ratios	**11:7**	**1:1**	**5:1**

Cite this article as:

Breit, F., Tamburelli, M., Gruffydd, I., & Brasca, L. (2024). Pushing boundaries in the measurement of language attitudes: Enhancing research practices with the L'ART Research Assistant app. *LingBaW. Linguistics Beyond and Within, 10*, 7–32.