Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages.

Pre-print of Chapter 1. Not intended for wider circulation.

Chapter 1: The IRIS Repository of Instruments for Research into Second Languages: Advancing Methodology and Practice

Emma Marsden (University of York); Alison Mackey (Georgetown University/University of Lancaster); Luke Plonsky (Northern Arizona University)

How we elicit our data is at the heart of empirical research. Debates around theories and constructs very often center around concerns about exactly how data were collected – Did the measure allow access to conscious reflection? Did the stimulus result in learners behaving realistically? Did any unintended factors impact our results and compromise our ability to explain them? Did we take into account the perspectives of targeted stakeholders? Improvements to how we collect our data can often lead to more persuasive tests of theory or constructs and/or more convincing applications in the real world.

However, tapping into language processing, learning, use and attitudes clearly poses varied challenges, particularly when the aim is to capture the nature and causes of change over time while using relatively static, 'one-off' windows into the phenomenon under investigation. Added to this are many other factors that can affect methodological choices, for example, the mode (input versus output), conscious awareness, age and cultural appropriateness, cognitive and perceptual systems all interact with how language is perceived and produced.

The current volume aims to address some of the issues involved in the collection of second language data. It emerged out of discussions at a conference "Eliciting data in second language research: Challenge and innovation", held in 2013 at the University of York. Each chapter builds on presentations and posters from that meeting[1]. The conference marked the culmination of the initial development phase for the digital database Instruments for Research into Second Languages (IRIS). IRIS is a unique, searchable repository containing materials used to collect data for studies reported in peer-reviewed publications in second-language

---

[1] with one additional, non-conference contribution (Ziegler).

(L2) research[2]. Each chapter in this volume reports on data collected with instruments available from the wide range of materials held on IRIS.

In this introduction to the volume, we have four goals. These are to:

1) Present the rationale, content, and usage of IRIS;

2) Provide summaries of the chapters in this volume;

3) Discuss themes emerging from the chapters;

4) Evaluate the role and challenges for IRIS in driving forward L2 research methodologies.

Part 1. Instruments for Research into Second Languages (IRIS): A digital repository of research materials

1.1 The rationale for IRIS

Researchers, funders, and research organizations are increasingly sharing *data* and *research reports*, and encouraging others to do so as well (see the Center for Open Science cos.org; r3data.org; dataverse.org; RCUK 2013; and learner corpora, e.g. Thomas (this volume)). However, the ways in which these data are collected is rarely fully described in most reports. That is, the full set of *actual* stimuli (e.g. pictures, participant instructions, response options, software setup, nature of distractors, test items) used to elicit the data is not usually provided. Even when authors would be willing to include the materials, stimuli, and so forth in the reports of their research, they are often limited by publishers' space constraints. Making these materials accessible provides two types of benefits. First, more thorough reporting improves the ability of consumers of research to evaluate the validity of the research. Second, access to the materials used facilitates the ease, speed and accuracy of all types of replication research, ranging from conceptual to exact (Marsden & Mackey, 2014; Handley & Marsden 2014; Mackey 2015). Indeed, in the second language research field, there are many reviews lamenting the paucity of studies that can be usefully compared even within a single agenda or

domain of research, given the range of factors that change between studies (e.g. outcome measures, intervention materials, linguistic feature) (Norris & Ortega 2000 Plonsky & Oswald, 2014). Increasing the comparability of sources of data, including their psychometric properties (e.g., Plonsky & Derrick, under review) will increase the systematicity and efficiency of our combined efforts. Calls for increased replication have been prevalent in many fields for decades (e.g. Amir & Sharon 1990; Freese 2007; Schmidt 2009), and, for research areas in which replication is deemed possible and useful, in our own field (Cumming, 2012; Porte 2012; Polio & Gass 1997; Santos, 1989).

In response to the need for and interest in greater transparency, some journals have relatively recently begun to hold supplementary materials for their publications online, linked to the specific article, journal, and publisher, and often behind a paywall. Other, smaller open access collections are available, but have specific focuses (e.g. Max Planck Institute http://fieldmanuals.mpi.nl/ with an L1 and cross-linguistic focus) or the collection at the University of Essex (http://experimentalfieldlinguistics.wordpress.com/experimental-materials/) with emphasis on collecting L1 usage data in the field. One of the main benefits of IRIS is that it is free and fully searchable across many parameters, for example, from 'research area' to 'journal', from 'funder' to 'participants' first language'. As IRIS is independent of any particular research group, country, funder or publisher, it has a wide reach, both geographically and in terms of its content.

1.2 IRIS content to date.

IRIS currently holds 1500 files, bundled into 590 complete sets of data collection tools.

One of the most striking features of IRIS, as just noted, is its 'searchability' across a large range of parameters, making it a unique contribution to open science. To make instruments searchable, they are each assigned descriptors from multi-levelled lists of terms to describe the nature of the instruments and how they have been used to date. Five of these are obligatory fields to be populated on upload: 'Reference to published study', 'research area', 'instrument type', 'file type' (image, audio, text), 'author/submitter of instrument'. The main IRIS ontologies are illustrated in summary form in Table 1. A total of 24 parameters are available to filter searches. Critically, the ontologies are organic, allowing uploaders to suggest '*other*' terms, which are then screened by the IRIS directors who accept suggestions as long as they do not paraphrase existing terms (and thus duplicate them). The ontologies

themselves are fully open access, via the 'Search Help' page, where each term itself is clickable, linking to all materials tagged with that term.

Parameter	Number of main terms (2nd ; 3rd level subterms)	Examples - each word searchable
Research area*	53	cross-linguistic influence; cultural identity; implicit learning; learner silence; pragmatic competence; prosody
Instrument type*	45 (55; 7)	communicative task; interview protocol; observation protocol; self-efficacy questionnaire; working memory test (reading span); stimuli for laboratory studies
Language feature being investigated	14 (41)	animacy; figurative language; lexis (collocations); morphosyntax (determiners); pragmatics (complaints)
Data type	11	reaction times; open response
Participant type	18	adolescent learners; artificial language learners; bilinguals; first language attritors; teacher trainees
Proficiency of learners	15 (28; 22)	complete beginner; CFER (waystage); ACTFL (advanced); TOEFL (speaking (fair))
Domain of language use	10	residence abroad; school (language instruction only)
Funder	20	European Commission; Fulbright
Language: 1st/2nd/3rd/4th		standard list of world languages
Authors cited*	602	
Journals*	115	

* Obligatory fields for uploaders to complete

Table 1. Ontologies for research into second languages: Examples of the IRIS metadata

The largest number of data collection tools are in the areas of 'Acquisition' (255); 'Morphosyntax' (220); 'Lexicon' (105); 'Effectiveness of teaching' (79); 'Processing' (58); 'Phonology' (42); 'Cross-linguistic influence' (41); 'Learner attitudes' (40); 'Interaction'

(39); 'Motivation' (33); 'Fluency' (32). (Clearly there is some overlap between these research areas, an inevitability resulting from intersecting research domains and evolving ontologies). In terms of quantity of materials, there are more grammaticality judgement tests (59); attitude questionnaires (57); instructional materials (51); gapfill / cloze tests (47); language background questionnaires (44); and fewer instruments like elicited imitation/sentence repetitions (25); stimuli for laboratory experiments (24); communicative tasks (23); interview schedules (21). IRIS also contains several working memory tests and interaction protocols, along with a wide range of other instruments. These types of tools are very representative of what is commonly used in second language research (e.g., Cohen & Macaro, 2010; Gass, 2009). However, given how large the field is and how many different sub-fields of interest exist and are emerging, the research areas and instrument types currently available through IRIS are not likely to be fully representative of the variety in the field as a whole. The main research interests of the two project directors who established IRIS, as well as other supporters such as the IRIS advisory board members, are likely to have, to some extent, (unintentionally) had an effect on the types of instruments initially uploaded, given their personal networks and fields of influence. Another factor contributing to what sort of instruments are contained in IRIS is the fact that some perspectives lend themselves to a number of relatively straightforward instrument types (e.g., GJTs, pictures for oral production), making file storage and access easy. In contrast, other research areas and approaches, including discourse analysis and pragmatics for example, tend to use (more) naturalistic data where the elicitation mechanisms are less clear, sharable, or indeed critical to the problem being addressed. Nevertheless, from the outset of the creation of IRIS, all team members have aimed for the database to be as widely inclusive as possible, and we continue to strive for contributions from a diverse range of theoretical and methodological traditions. The growth of IRIS depends on the community sharing its instruments, as well as using the ones already made available.

Of course, some researchers also make their materials available on their own personal or project websites. With the researchers' permission, IRIS also hosts these, bringing advantages for both the researchers and the repository. IRIS links to the project/personal websites, but also holds the material and their metadata, making them searchable and increasing their visibility, e.g. Paul Meara's vocabulary research materials (from http://www.lognostics.co.uk/); Monika Schmidt's attrition research website (from http://www.let.rug.nl/languageattrition/); Zoltán Dornyei's motivation research tools (from

http://www.zoltandornyei.co.uk/); Georgia Institute of Technology's attention and working memory lab's tests (metadata only held in IRIS, http://englelab.gatech.edu/tasks.html).

IRIS also holds an increasing amount of metadata for research materials that are not actually held on the repository, thus enhancing the completeness of IRIS as a database of methodologies. Metadata are routinely entered on a weekly basis for all research materials used for empirical studies in issues of major journals that regularly publish research into second language learning and teaching (e.g. *Applied Linguistics, Applied Psycholinguistics, Bilingualism, Language & Cognition; Language Learning; Linguistic Approaches to Bilingualism*; *Modern Language Journal*; *Second Language Research; Studies in Second Language Acquisition; TESOL Quarterly*). Once metadata are uploaded, authors are then invited to contribute their full materials.

Two additional pathways by which researchers—whether as contributors, browsers, or downloaders of IRIS materials—become familiar with IRIS is through applied linguistics journals. The field's journal editors have been extraordinarily supportive of the IRIS initiative and the move it represents toward great transparency and research that is more community-based rather than individualistic. These editors of journals (e.g., *Applied Linguistics; Applied Psycholinguistics; Language Learning; Linguistic Approaches to Bilingualism; Second Language Research; Studies in Second Language Acquisition; The Modern Language Journal; TESOL Quarterly*), and their publishers, actively support IRIS through the journals by recommending that data collection materials (e.g. intervention tasks, outcome measures, stimuli) be submitted to IRIS, via the editor's notification of acceptance for publication and/or via the guidelines for authors. See Appendix A and the IRIS website for the list of over thirty journals that support IRIS in this way.

In other words, IRIS provides a large scale reflection of methodologies used in publications in second language research. With long-term funding from the *British Academy*[3], and as a long-term curation of the University of York Digital Library, IRIS is set to be a sustainable resource for the community, capable of responding to its needs, as well as the demands of ever-changing digital technology.

---

[3] IRIS is an '*Academy Research Project*' AN110002, 2012-17, with anticipated renewal of funding in subsequent five year tranches.

1.3 How IRIS is being used

The total number of downloads in its first two and half years of existence has now reached 8000, with every other visit to the site resulting in a download (i.e. 16000+ unique visits, approximately 18 visits per day). Language teachers, tutors and teacher educators comprise approximately one fifth of the downloaders, with students representing three fifths and academics the remaining fifth of downloads, including both early and well established scholars (see Figure 1). Users have the facility to 'Request' materials by submitting the reference details of the materials they would like to see. The IRIS team then contacts the researchers who published using those materials, inviting them to contribute their materials, pointing out the benefits that this would have. To date, approximately a third of these invitations results in authors contributing their data collection tools, with encouraging indications that this rate is increasing.
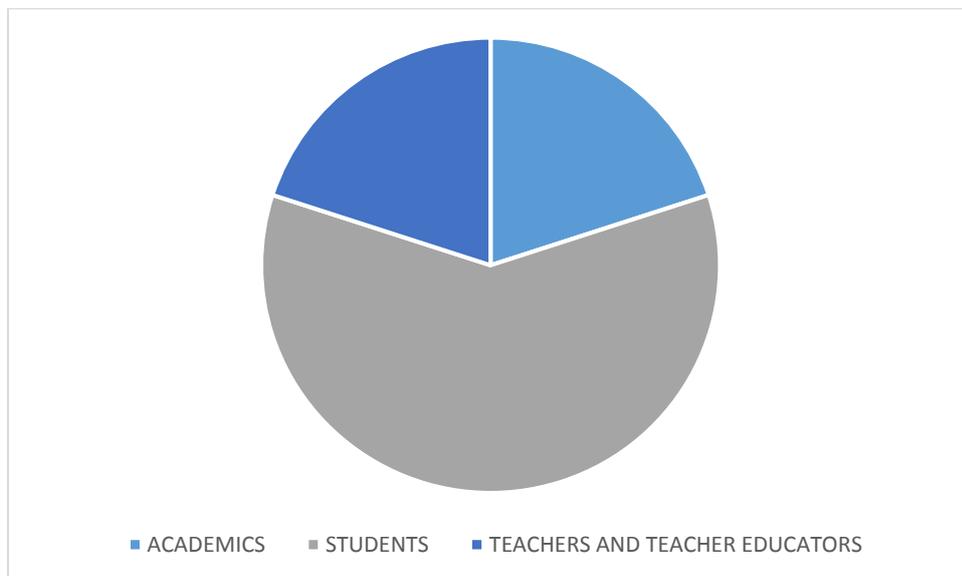
INSERT FIGURE 1 HERE

FIGURE 1: Proportions of downloads from IRIS by downloader's status (total downloads = 8000 at time of going to press)

Although the criteria for inclusion on IRIS is that the materials have been published in peer-reviewed journals, conference proceedings and books, the vast majority of materials on IRIS is linked to articles in journals (approximately 500 of the 590 current instruments). Downloads are gradually aligning with journal citation indices, with more downloads of

materials that have been used for publication in highly cited journals. This is some indication that IRIS and journal publications are likely to reinforce reciprocal visibility. Clearly, the number of materials uploaded that are linked to specific journals also tends to increase downloads of materials linked to those particular journals, a benefit which is probably also appreciated by supporting editors. Download and upload information is also freely available using the 'Statistics' tab on IRIS, making usage information available for the community. In addition, contributors can be notified about downloads of their instruments via email or on their personal 'My IRIS' page.

Part 2) The chapters in this volume.

The first two sections of this volume (chapters 2-10) present critical reviews of particular methodologies that include detailed evaluations of empirical studies. Common to all chapters is a heightened concern for the validity underlying various data collection techniques. All of these chapters also investigate the oral modality, with two exceptions that review studies that have largely used written stimuli (Williams & Pacriorek, and Roberts). The first group of chapters (2-5) all focus on learners' responses when exposed to input, which is widely accepted to be a critical phase in learning. Methods discussed in this section have the advantage of being able to target particular language features in the stimuli, allowing investigation of features that can be challenging to elicit in production, either because learners can avoid using the features, and/or become highly aware of them, and/or because (lack of) prior knowledge interferes with production. The second group of chapters (6-10) focus on the elicitation of oral production. Methods discussed in this section have the advantage of tapping into a highly externally valid construct: the ability to produce the language, often under communicative pressure. The final three chapters (10-12) provide broader methodological reviews or position papers.

2.1 Tapping into the processing of input: Chapter 2-5

These chapters each adopt a different approach to investigating what learners do when exposed to input, many of them discussing the nature of knowledge learners tap into when undertaking the task. We present the chapters in an order that indicates decreasing concern about avoiding conscious knowledge of learning the target structure: The first chapter (Williams & Pacriorek) addresses how to reduce, as far as possible, tapping into learners' access to awareness of the target structure during data collection; in contrast, the last chapter (Godfroid & Spino) sets out to intentionally tap into such awareness. Those in between have

varying degrees of concern over this matter, some are more concerned with exploring external validity and practicality of measuring the effects language exposure has on behaviour.

Williams & Paciorek, whilst acknowledging that no test can guarantee the *eradication* of explicit knowledge, give a thorough account of 'indirect tests', techniques that reduce conflation with learners' use of conscious knowledge at the point of testing. To construct tests that maximise access to implicit knowledge, their studies make no distinction between learning and testing phases, in which the training provides evidence of a grammatical rule and the testing phase introduces violations of that rule. Whilst learners' conscious attention is intentionally focussed on one task, their reaction times are providing a window into their non-conscious knowledge. This paradigm rests on the assumption that slower reaction times reflect some level of knowledge or sensitivity that a grammatical rule is being violated. Williams & Paciorek also stress the importance of combining such designs with knowledge source judgements, in which participants indicate the basis of their response (e.g. guess, intuition, rule, memory).

Lew-Williams provides a thorough overview of the technique 'looking while listening' – a practical type of visual world eye-tracking providing a window into how oral input is processed for meaning in real time. Not only does the chapter contain many practical tips, it also argues that eye movements can provide valid evidence about the relationship between online processing efficiency and learning. Lew-Williams gives a critical review of different types of evidence about efficiency of processing, including reaction times, accuracy (of looking to the target referent), and shift-tendency (a count of gaze changes between the target and distractor referent(s)). He also draws on important theoretical debates, about the relationships between *processing, learning* and *proficiency*, usefully citing several studies that have found relationships between processing efficiency and measures of productive language L1 proficiency. To illustrate the potential contribution of the technique, Lew-Williams reviews several investigations that have looked at the processing of gender and number in Spanish L2. The critical manipulation in these studies is whether or not a morphosyntactic marker (e.g. gender or number) disambiguates which referent is being talked about *before* the noun itself is heard, evidenced by faster reactions for looking at the target referent. Lew-Williams and colleagues have found that whilst adult L2ers were able to use abstract morphology for predicting biological gender and number, they did not process markers for grammatical gender. The review provides an excellent example of how

instrumental and conceptual replication can provide many insights over a series of studies, into, for example, the effects of age, L1/L2, proficiency, and abstract linguistic versus conceptual categories.

Roberts provides a state of the art review of the use of self-paced reading to investigate sentence processing, describing the incremental, millisecond parsing decisions made by learners as they read sentences word by word. Roberts provides many examples of studies that remove semantic biases (or anticipated meanings that are induced due to expected collocations or properties of the lexicon) in the input that, ordinarily, would help learners to assign meaning to the input. However, once these biases (or induced expectations) are removed, learners, unlike natives, cannot use other cues (such as case or number marking, to reassign a meaning during real time. Roberts stresses that task instruction appears to influence findings, as asking learners to perform a task whilst listening, such as judging grammaticality, may alter processing behaviour. She also stresses the need to take into account learner differences such as the L1, WM and proficiency.

Godfroid & Spino describe a novel technique for eliciting 'noticing' known as 'finger-tracking'. They carried out two types of triangulation in order to explore the usefulness of this technique, in the context of a study looking at nonce vocabulary learning. In their study, L2 English learners were asked to read whilst tracking the words with their fingers as well as think aloud. Godfroid and Spino found that a subjective dichotomous decision about whether the finger had paused (+/- pausing) had a strong positive correlation with millisecond timings of the time the finger paused. They also found that, although there was a large degree of overlap between finger pauses and thinking out loud about the novel words, finger pausing was more frequent than thinking aloud, and accounted for an additional 13% of instances of noticing the target feature. Finally, Godfroid & Spino note that although +/- pausing did not make additional contribution to predicting vocabulary recognition, the best predictor included the millisecond timings of finger pausing combined with the presence/absence of thinking aloud. They conclude by discussing potential advantages of finger tracking over eyetracking when these techniques are used with concurrent think-alouds, not only in terms of expense and ease, but also in terms of construct validity. They speculate that length of dwell time that is not affected by eyes 'wandering', which is a problem for eye-tracking, may provide a valid indication of depth of processing.

3.2: Tapping into production: Chapters 6-10

The following five chapters all address the elicitation of oral production. The order of the chapters arguably represents 'more' to 'less' controlled elicitation methods, with chapter 6's technique providing complete control over stimuli and the intended productions, and chapter 10 using a corpus of oral production data which exerted little control over the features that were elicited.

Erlam & Akakura provide a state of the art review of elicited imitation, a technique that is increasingly popular as it allows researchers to 'trap' specific forms under investigation, whilst also eliciting oral production. By considering different facets of validity (construct, criteria, predictive and response) they evaluate the likelihood that elicited imitation tests are tapping into language knowledge, rather than rote memory. They then present an innovative adaptation of elicited imitation, 'elicited production', which uses a story and comprehension questions to increase learners' focus on meaning. The authors demonstrate the potential importance of collecting evidence that learners correct ungrammatical sentences, and conclude by setting out an agenda for a series of replication studies, including the need for studies that combine measures of awareness with elicited imitation studies to assess the extent that they can be viewed as tapping, to some extent, into implicit knowledge.

McDonough, Kielstra, Crowther & Smith present an original laboratory study with learners of English with a variety of L1s, adopting one of the few 'pre-post' designs in priming literature. They tested whether enriching an interlocutor's input with exemplars of a specific type of post-nominal modifier (relative clauses) increased learners' productions of this syntax. Results showed a large increase in relative clause use, which was greater when the previous utterance was a relative clause (primed contexts) than when it was not (unprimed). They investigated, in line with several other chapters in the volume, the extent to which awareness may have played a role in the observed performance, as well as the extent to which WM (backward digit span) and statistical learning (pattern inferencing) may relate to priming effects. They found descriptive evidence that retrospective awareness of the target structure seemed to lead to a slightly increased amount of production, but this was not statistically robust. Although WM and statistical learning measures correlated with 'awareness', they did not relate to actual production of relative clauses. The authors consider the influence of proficiency level in this finding and the extent to which priming may influence existing representations, rather than scripted priming interaction activities establishing new ones.

Garcia-Mayo & Alcón provide a detailed overview of a range of studies that have used two task types to elicit production: dictogloss and discourse completion. They describe how dictogloss tasks provided opportunities for Language Related Episodes (LREs) in which university English as a Foreign Language learners talked about language whilst having to co-produce it. The LREs included talking about grammar, and this related to accurate use in the final version the learners produced. Garcia-Mayo & Alcón provide evidence that asking learners to write, rather than say, their final version leads to more LREs, more focus on grammar and more accurately resolved LREs, with some, but relatively little, use of the L1. They note that the extent to which such tasks are useful for younger learners requires conceptual replications. This chapter also critically evaluates the extent to which discourse completion tasks (DCTs) provide a window into how learners perform speech acts and how this can be contextually bound. Acknowledging their weakness as eliciting speech in the written modality, they track the relatively recent innovation of aural-oral DCTs. Such tasks have also been used as prompts for co-production to study LREs, in peer and teacher-student interaction, and as scenarios for role plays that elicit and develop pragmatic ability. By using retrospective verbal reports, the authors describe how Alcón investigated the thoughts that bilinguals could articulate whilst completing DCTs eliciting refusals. In line with many chapters in the volume, the authors are concerned that future research investigate the nature of knowledge that is used (explicit/implicit), particularly in more natural classroom and communicative environments as opposed to exclusively in laboratory contexts.

Philp discusses three challenges of using tasks to elicit oral production data between second language learners. Two challenges that she addresses concern the quality of elicitation, specifically, authenticity and creative use of the target language, while the third challenge concerns learner engagement. Philp specifically reviews two data sets (Oliver, Philp & Mackey, 2008; Philp, Oliver & Mackey, 2006; and Philp & Iwashita, 2013) used for oral production elicitation in order to discuss each concern. She uses multiple examples of learner-learner interaction from the data to illustrate that a sole focus on just one of the three challenges when selecting tasks cannot be rendered sufficient. She also highlights the importance of factors such as age, setting, and purpose of the task being used. By using specific examples, Philp successfully demonstrates the need for careful consideration of task selection specifically when they are used to elicit learners' oral production. Overall, she advocates for a three-way balance within tasks, in that they elicit the target form in an authentic context, enhancing the ecological validity of the research, and in ways that are

linguistically and cognitively challenging for learners, ultimately ensuring all three: authenticity, creativity, and engagement.

Thomas also provides an original empirical study, as well as an overview of research to date on syntactic priming. Thomas introduces several innovations: she uses the somewhat neglected method of analysing corpus data to investigate the role of input and interaction (normally investigated using tightly controlled laboratory experiments); she focusses on French as an L2; and on a morphosyntactic feature (*être* auxiliary use in French perfect tense) which is constrained by both syntax and semantics. She provides a fine-grained, learner-level analysis of the use of 'naturalistic' priming of this feature over 2-3 years, amongst Swedish immersion learners of French. Although the tokens are few in number, the analysis provided apparent evidence that learners' usage of the *être* auxiliary was greater in interactions in which the learners' interlocutors used it compared to when they didn't. She also found that this did not seem to be merely a function of the precise form of *être* used by the interlocutor. Although not setting out to test interactionist theories, Thomas' evidence aligns with such perspectives. She also discusses saliency, cross-linguistic influence, and mutual alignment as potential factors in the extent of priming effects, again, all inviting further research.

3.3 Research design & context

Ziegler presents an account of the designs and reporting practices in primary research that has examined the relative effects of interaction in Synchronous Computer Mediated Communication and Face to Face environments. Her study is informed by and speaks to two unique sets of researchers in applied linguistics. The first set is comprised of researchers in computer-assisted language learning (CALL) who have long since been calling for improved theoretical and methodological rigor in their own domain (e.g., Chapelle, 2001; Macaro, Handley, & Walter, 2012). In the other set are those who have promoted research synthetic techniques not only as a means to bring together substantive effects in a given domain but to describe and evaluate the methodological practices leading to those effects as well (e.g., Norris & Ortega, 2000; Plonsky, 2013; for a review of such efforts, see Plonsky & Gonulal, 2015). Ziegler's methodological synthesis includes a total of 16 primary studies, all of which were coded for features related to study designs, analyses, and reporting practices. Her results reveal a number of inconsistencies such as a lack of delayed posttests, effect sizes, confidence intervals, and checking of statistical assumptions. Critically, the chapter does not end with a laundry list of methodological grievances as is often the case with this sort of research. It

provides instead a careful set of forward-minded suggestions for how research in this domain can improve in the hopes of producing more valid and reliable findings, enhanced theoretical models, and improved research practices. We also note that her data collection instrument, like many others used and discussed in this volume, could certainly be modified and applied in order to gather data on the methodological practices of other domains of L2 research.

Oliver and Grote provide a description of the research methodology they have used to develop a Task-Based Needs Analysis (TBNA) model for a high school specializing in Vocational Education and Training (VET) for Aboriginal learners who speak Standard Australian English as an additional language. They illustrate how they use a systematic, evidence-based Needs Analysis (NA) to identify the workplace language and literacy needs of learners. In other words, they extended the methodology of traditional NA in a unique way, demonstrating the importance of taking into account the perspective of all major stakeholders, not only inside the future workplace of the learners but also inside their cultural domain. Oliver and Grote argue that, in the context in which they conduct research, traditional NA often fails to address less visible cultural issues between learners and future employers. As a result, a traditional NA-based curriculum prepares learners less effectively for employment than a TBNA one. While their methodological procedures equal the traditional NA approach (interviews, observations), they present a strong case that the range of data sources should be extended to stakeholders in the cultural domain. In their chapter, Oliver and Grote specifically present analyses of interviews with (1) workplace domain experts (VET teaching staff and employers) and (2) Aboriginal cultural insiders. By comparing and contrasting the results these two data sets, they are able to illustrate the need for both perspectives when developing a TBNA model.

Al Khalil also raises research design issues, first reviewing different approaches for tapping into state motivation, as it ebbs and flows in relation to specific learning tasks, in order to provide 'person-in-context' accounts of learner motivation. She then presents original data collected from a 'motivation thermometer', completed by learners immediately after they completed oral interaction tasks which, four months previously, they had rated in terms of their importance as learning activities. Little correlation between the prior ratings of 'importance' and post-task ratings of motivation provides a neat demonstration of the need to investigate the motivational value of tasks *in situ* with learners undertaking example tasks. Al Khalil also shows how widely individuals' profiles can vary, demonstrating the importance of analysis at participant, not just group, levels.

4   Key methodological challenges and innovations raised in the chapters

Key themes emerging from the chapters include the relationship between 'construct' and 'elicitation techniques/design', as well as the importance of considering context and participants.

4.1   Instruments and their construct validity

The choice of research instrument is driven by the constructs that researchers need to tap into. One theme evident in almost every chapter is the need for researchers to clarify the nature of knowledge that is accessed during any elicitation task, and the extent to which learners have access to conscious knowledge during learning and testing phases.

Models of language. Another, arguably related, theme is the nature of the language representations being accessed. For example, in terms of representations that can be primed, Thomas investigates the use of *être* auxiliary, and analyzes instances where an interlocutor's use of one inflection (e.g. *est*) can prime the use of another inflection (e.g. *suis*), thus assuming that learners have access to a store of irregular inflections that are connected in the mind in a paradigmatic, componential fashion, which can therefore prime each other. Thomas also suggests that this store is networked with particularly lexical items: verbs with specific semantic properties relating to telicity and unaccusivity. This assumes a componential view of how irregular inflectional morphology is stored and accessed, rather than an item-based or holistic view (see Marslen-Wilson 2007 for discussion).

Models of learning and priming. The research designs included in the volume reflect various concerns about whether the research aims to tap into *learning* mechanisms or *processing* behaviors.   For example, some studies set out to monitor constructs (whether these are learning effects, in terms of speed or accuracy of access, or motivation or attitudes) during some activity or training over time (whether a few minutes or years) (this volume, Williams & Pacriorek; McDonough et al.; Garcia-Mayo & Alcón; Philp; Thomas; Al Khalil; Oliver), whilst other chapters focus on documenting online behaviour at a particular stage of development (this volume, Roberts; Lew-Williams). This raises several challenges in terms of how these two areas of research may interface: online processing and learning. One challenge for research is to investigate the extent to which online processing influences and drives *learning* (e.g. O'Grady, 2015) or whether it serves, rather, to reflect current knowledge (which may, perhaps, have been learned by mechanisms other than processing itself).

Coding the data and construct validity. Of course, coding and analysis decisions are equally as critical for the interpretation of findings as the instruments themselves. For example, McDonough et al. identified priming effects as those in which the prime occurred in the immediately preceding utterance, including both the interlocutor's and learner's own primes. Thomas, on the other hand, considered syntactic priming to have occurred if a target was produced in the same 20-30 minute recording, and only if produced by the interlocutor (i.e. between-, not within- speaker, primes). As Thomas points out, both these choices carry some element of risk of misrepresenting the construct, as 'immediate-only' tokens could reflect echoic repetition (rather than the priming of syntactic representations) or provide an overly conservative estimation of priming effects, and long timelags could reflect 'independent' (unprimed) productions. Repetitions and independent productions are difficult to identify if analyses include both same and different lexical exemplars, in prime-target pairs, of the target phenomenon. Thomas' analyses of items and turn-counts provide critical insight into these issues, and both sets of authors invite conceptual replications to tease out such issues.

4.2   External and ecological validity: Context, participants, languages, and features

Participants. In line with much SLA research, a number of the studies presented or reviewed in this volume have worked with participants who are young adult/teenage learners of English or major Western *foreign* languages (McDonough, Garcia-Mayo & Alcon, Thomas, Lew-Williams, Al Khalil). However, both Oliver and Philp emphasise the need to improve ecological validity, by involving a wider range of participants and seeking their views. Only by studying wider populations, better reflecting the variety of contexts for millions of learners worldwide, will we be able to develop theoretical models robust enough to account for all learners (e.g., Spinner, 2011).

Languages and language feature. Many SLA studies try to isolate effects of the L1, and so their participants tend to have a specific L1 (or set of L1s) or limited or no other language learning experience (e.g. in this volume, Al-Khalil, Garcia-Mayo & Alcón, Lew-Williams, Roberts, Thomas, Williams & Paciorek). This has clear ecological validity, given that many people learn a foreign language along with (a majority of) others in instructed settings who share the same L1. On the other hand, researchers also include samples of learners with a range of prior language experience (e.g. in this volume McDonough at al., Philp, Oliver), which is also ecologically valid, given that in many ESL contexts multilingual experience is the norm.

In terms of language features under investigation, most of the chapters in this volume focus on morphosyntax, with some studying vocabulary. This focus and balance is in line with much SLA work, and is largely reflected in IRIS where 254 instruments are tagged with 'morphosyntax' and 150 with 'vocabulary' as we noted earlier. However, IRIS also hosts 66 instruments for researching 'phonology'; 19 for 'pragmatics'; four for 'figurative language'. It is a key aspiration to ensure that the materials on IRIS represent the full breadth of SLA work.

5   Conclusions: Challenges and avenues ahead for IRIS

5.1   Current and future directions of the open access movement.

The open access agenda is not without its critics (Osborne, 2013; and for debate see Vincent & Wickham 2013), as well as its silent abstainers. One concern, for example, is intellectual property. In line with most open access resources, in IRIS contributors of materials have to assign a *Creative Commons License* to their material. Downloading any material signals agreement to cite the reference(s) attached to that instrument. Given that plagiarism detection software, which harvests data available on the internet, is now regularly used by publishers and universities, having one's materials on IRIS means that misusing them should be detectable. Also of relevance to this kind of concern is that improper citing of materials is concomitant with research quality issues. Therefore, it is, of course, still incumbent upon our community to ensure that use or adaptation of materials held on IRIS gives due consideration to quality and appropriateness (as we also discuss below in section 5.3).

Indeed, the open access movement does not seem hindered by concerns about intellectual property, and looks set to continue with its impact reaching many publishing houses. Publicly funded bodies are increasingly insisting that products of publicly funded research are made freely and openly available (RCUK, 2013; http://roarmap.eprints.org/). Endeavors such as the Centre for Open Science (COS) provide a platform (the Open Science Framework) for large multi-site projects, facilitating the sharing of protocols, sample sizes, power analyses, data etc. Furthermore, COS, in collaboration with journals, now awards 'badges' to recognize and promote open science practice. IRIS is now one of COS's recognised repositories. Studies that have stored either their data, materials, or pre-registered their protocol on an open access forum can display the relevant open access symbol on the published article (the field of applied linguistics is adopting these standards, as illustrated by Trofimovich, 2015 where the journal *Language Learning's* policy is outlined).

There is also evidence that open access practices improve science and its dissemination in several ways. For example, holding open access *data* has been associated with better research and reporting quality (Wicherts, Bakker & Molenaar, 2011) and with increased citations (Piwowar & Vision, 2013); publishing open access *reports* can increase citations (Wagner, 2010). Whether similar benefits result from research associated with open access *data collection materials* remains to be investigated, and IRIS provides a fertile ground for exploring this possibility.

5.2 Replication

As we noted earlier, one of the main aims of IRIS is to promote replication in response to repeated calls for more replication in L2 research given that it can provide "essential, methodical support for theory" (Porte, 2012). Our field has seen several discussion pieces promoting and discussing replication (e.g. Porte & Richards 2012, Mackey, 2015), and many calls exist for increased replication (e.g. Handley, 2014; Handley & Marsden 2014), via, for example, students carrying out confirmatory research (Polio & Gass 2007; Porte 2012 & 2015). However, although special sections are available for the publication of sound replications in a few journals (e.g. *Language Teaching; Studies in Second Language Acquisition*), these have not been frequently filled. Several reasons have been suggested for the small quantity of published replication work, including poor incentivization in funding, publishing, and promotion systems, resulting in low perceived worth and esteem of replication research.

One way of demonstrating the need for replication is the extent to which replications do or do not actually successfully replicate the findings of the original study. To this end, Marsden & Mackey (2013) searched for all studies that self-labelled as replication studies, having the term 'replicat*' in their titles or abstracts. The investigation surveyed all journals in the database *Language and Linguistic Behaviour Abstracts* between 1973 and 2013, finding just 40 studies that self-labelled as 'replication'. Of these 40, 32 could be assessed for the extent to which they supported, or otherwise, the original study. 21/32 supported or corroborated original findings; 13/32 found context dependent differences; 12/32 produced contradictory findings. Of course, it is difficult to distinguish between studies that successfully replicate findings and those that find context-dependent differences, as well as between studies that find context-dependent differences and those that find contradictory findings. But, broadly speaking, these findings suggest that without replication, findings from

individual studies can produce misleading or incomplete findings (see also Klein et al. 2014; the Reproducibility project https://osf.io/ezcuj/). Marsden & Mackey (2013) also pointed to a relatively narrow focus in self-labelled replications (with an emphasis on morphosyntax, processing) that used particular tools (a majority using writing tests and questionnaires). This approach contrasts with the broader view of replication often embodied by meta-analysis wherein all studies being synthesized might be considered conceptual replications (e.g., Plonsky, 2012). Indeed, Williams & Paciorek (this volume) demonstrate the challenges and importance of replication through their discussion of what might appear to be different findings from studies by Leung & Williams (2012 & 2014) and Leow & Hama (2013). Some of the differences were due to different operationalizations of materials and design. McDonough et al. (this volume) also present an empirical study that offers many opportunities for theoretically-driven replication, such as investigating the effects of the L1, the nature of the target feature (e.g. object *versus* subject relative clauses; or less *versus* more marked target features), or different measures of WM.

One suggestion to increase the amount of published replication/confirmatory research is that replication should be carried out by the original research team and published within the original article (e.g. Porte, 2015; VanPatten, 2015). Whilst this may be an ideal goal, its feasibility (e.g. cost and time) is unclear, and it would require publishers to provide more space per article. More importantly, one of the strengths of *multi-site* replication (via, for example, pre-registered protocols for external research teams to follow) is that it reduces the likelihood of any biases that may be held by the research teams and provides a more rigorous testing ground for generalizability and reliability. For example, Makel, Plucker & Hegarty (2012) found that in psychology research, when the original authors were involved in the replication study, the odds that the original findings would be replicated was 92%, compared to just 65% when undertaken by independent researchers. See also Makel & Plucker (2014) for similar findings in educational research.

Making an initial step to try to address some (by no means all) of these concerns, Marsden & Morgan-Short (ongoing) are coordinating a multi-site replication study in SLA, with both invited multi-site replications and an open call for multi-site collaboration. A pre-registered protocol is posted on the Center for Open Science's Open Science Framework and all data collection materials for the study are contained in IRIS.

Clearly, IRIS has great potential to facilitate replication, particularly instrumental and operational replications, because it can contain coding/scoring information that is essential for robust replications. The materials in IRIS can also facilitate other types of replication too, including, for example, conceptual replications with different dependent and independent variables in different contexts, L1-L2s, ages, and so on. The extent to which this opportunity is taken up by the field will become clear in the long term. Significant use of IRIS for replication will probably also require more cultural change in the way the field views replication, in terms of mechanisms such as funding, publication, tenure, as discussed above.

5.3 Harnessing the potential of IRIS for research methods training: Engaging with early career researchers and language education practitioners.

IRIS does not provide judgements about quality, above and beyond quality assurance based on the standards of the field. In other words, quality is promoted by ensuring the instruments have been used in publications that have been peer reviewed or given institutional approval (for PhD dissertations). IRIS is not intended to lead to 'off-the-shelf' use of materials by researchers who have not carefully considered the suitability of the materials for their own research purposes and design. Obviously, not all materials are axiomatic exemplars of specific techniques. Encouragingly, the (admittedly quite limited) feedback provided so far by downloaders suggests a range of uses: from looking at the materials and considering them carefully for use or adaptation as part of designing their own work, to using them for teaching and advising purposes, thus drawing on IRIS as a resource that facilitates and enriches critical thinking and skills development.

As we said earlier, one of the rationales for IRIS is to increase the transparency of research for practitioners (e.g. language tutors, teacher educators and policy-makers), both as consumers of research and as researchers undertaking their own context-dependent research (Borg 2013; Burns 2010). To promote this goal, workshops, presentations, and publications for practitioners are frequently given by the IRIS team and colleagues (e.g. King & Marsden, 2013)[4]. For practitioner researchers, with limited research expertise, time and funding to develop their own materials, IRIS has some potential for increasing their research awareness and repertoire. For example, cost and practicality of doing research are important considerations, yet are often perceived as 'taboo'. In this volume, Lew-Williams and Godfroid openly tackle these concerns in their accounts of 'looking while listening' and

---

[4] For full list of practitioner-oriented activities, see http://www.iris-database.org/iris/app/home/outputs

'finger tracking' techniques, which increase options for those who do not have access to expensive equipment. Similarly, several studies emphasized the need for individual-level analyses (this volume, Thomas, Al-Khalil, Oliver). This need has positive practical implications: expensive, large scale experiments and surveys are not always necessary and may, in some cases, actually obscure interesting developmental trajectories.

5.4 Content of IRIS

By working with journal and book editors and conference organizations, the scope of the content in IRIS is steadily broadening. Recent inclusions, for example, are materials used to investigate artificial language learning, bilingualism, figurative language, and identity. However, as discussed earlier, there is still low representation of particular types of materials and research areas, such as neurolinguistics, emergentist and usage-based perspectives, L2 education policy, and task-based teaching intervention activities.

In this volume we have collected chapters to illustrate some of the *breadth* of theories, contexts, research areas, and methodologies hosted within IRIS. We hope that IRIS also promotes greater *depth* of research, by increasing the transparency and replicability of individual studies. We are grateful to all the researchers and users who are helping to enrich and sustain IRIS, providing an evolving resource and an archive of the elicitation methods in our field for future generations.

Appendix A: Journals that promote IRIS to their contributing authors, via acceptance notes from the Editors and/or via author guidelines.

Annual Review of Applied Linguistics
Applied Linguistics
Applied Psycholinguistics
The Asian EFL Journal
Assessing Writing
Bilingualism: Language and Cognition
Canadian Modern Language Review
ELT Journal
English for Specific Purposes
Hispania
International Journal of Bilingualism
International Journal of Bilingual Education and Bilingualism
International Review of Applied Linguistics
Journal of English for Academic Purposes

Journal of French Language Studies
Journal of Pragmatics
Journal of Research in Reading
Journal of Second Language Writing
Language, Culture and Curriculum
Language Learning*
Language Teaching
Language Teaching Research
Language Testing
Lingua
Linguistic Approaches to Bilingualism
Linguistics and Education
The Modern Language Journal
Second Language Research
Studies in Second Language Acquisition
System
TESOL Quarterly

References

Amir, Y., Sharon, I. (1990) Replication research: A "Must" for the Scientific Advancement of Psychology. *Journal of Social Behavior and Personality* 5 (4), 51-69.

Borg S. (2013) *Teacher research in language teaching: A critical analysis*. Cambridge, UK: Cambridge University Press.

Burns, A. (2010). *Doing Action Research in English Language Teaching: A Guide for Practitioners*. New York: Routledge.

Chapelle, C. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research*. Cambridge: Cambridge University Press.

Cohen, A. D., & Macaro, E. (2010). Research methods in second language acquisition. In E. Macaro (Ed.), *The Bloomsbury companion to second language acquisition* (pp. 107–133). New York, NY: Bloomsbury Academic.

Cumming, A. (2012). Comparative research, research syntheses, and adopting instruments in second language writing. *Journal of Second Language Writing*, 21 (3), 298-299

Darley, R. Reynolds, D. & Wickham, C. (2014) *Open access journals in humanities and social sciences*. London: British Academy.

Freese, J. (2007). Replication Standards for Quantitative Social Science Why Not Sociology? *Sociological Methods & Research*, 36 (2), 153-172.

Gass, S. (2009). A survey of SLA research. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 3–28). Bingley, UK: Emerald.

Handley, Z. (2014). Constructing an evidence-base for future CALL design with 'engineering power': The need for more basic research and instrumental replication. *EUROCALL Review*, 22 (2), 46-56.

Handley, Z., & Marsden, E. (2014). Linking CALL and SLA: Using the IRIS database to locate research instruments. In S. Jager, L. Bradley, E. J. Meima, & S. Thouësny (Eds), *CALL Design: Principles and Practice: Proceedings of the 2014 EUROCALL Conference*, Groningen, The Netherlands (pp. 1-6). Dublin: Research-publishing.net.

Klein, R. et al. (2014). Investigating variation in replicability: A ''Many Labs'' replication project. *Social Psychology* 45 (3), 142–152.

Leow, R. P., & Hama, M. (2013). Implicit learning in SLA and the issue of internal validity. *Studies in Second Language Acquisition, 35*, 545-557.

Leung, J. H. C., & Williams, J. N. (2012). Constraints on implicit Learning of grammatical form-meaning connections. *Language Learning, 62*, 634-662.

Leung, J. H. C., & Williams, J. N. (2014). Crosslinguistic Differences in Implicit Language Learning. *Studies in Second Language Acquisition, 36*, 733-755.

Macaro, E., Handley, Z., & Walter, C. (2012). A systematic review of CALL in English as a second language: Focus on primary and secondary education. *Language Teaching, 45*, 1-43.

Mackey, A. (2015). Methodological practice and progression in second language research. *AILA Review*, 80-97.

Makel, M. & Plucker, J (2014). Facts are more important than novelty: Replication in the Education sciences. *Educational Researcher,* 43 (6), 304 –316.

Makel, M., Plucker, J. & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science,* 7 (6) 537 –542.

Marsden, E. & King, J. (2013). The instruments for research into second languages (IRIS) digital repository. *The Language Teacher* 37 (2), 35-38

Marsden, E, & Mackey, A. (2013). IRIS and replication. Paper presented at the 9th *International Symposium on Bilingualism,* Singapore, 10-13 June 2013.

Marsden, E. & Mackey, A. (2014). IRIS: A new resource for second language research. *Linguistic Approaches to Bilingualism,* 4 (1), 125-130.

Marsden, E. & Morgan-Short, K. (ongoing). A multi-site replication study: Does attending to form affect comprehension? *Project funded by a Language Learning small grant.*

Marslen-Wilson W. D. (2007). Morphological processes in language comprehension. In G. Gaskell (Ed.) *Oxford Handbook of Psycholinguistics*, 175-193. Oxford: Oxford University Press.

Norris, J. and Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50 (3), 417–528.

Norris, J. and Ortega, L. (Eds.) (2006). *Synthesizing research on language learning and teaching.* Amsterdam: John Benjamins.

O'Grady, W. (2015). Processing determinism. *Language Learning* 65 (1), 6-32.

Osborne, R. (2013). Why open access makes no sense. In N. Vincent & C. Wickham (Eds.) *Debating Open Access*. (pp 96-105). London: The British Academy.

Plonsky, L. (2012). Replication, meta-analysis, and generalizability. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 116-132). New York: Cambridge University Press.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition,* 35, 655-687.

Plonsky, L., & Derrick, D. J. (under review). A meta-analysis of reliability coefficients in second language research. Manuscript under review.

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning,* 65, Supp. 1 (edited by J. M. Norris, S. J. Ross, & R. Schoonen), 9-36.

Polio, C., & Gass, S. (1997). Replication and reporting: A commentary. *Studies in Second Language Acquisition, 19* (4), 499-508.

Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878-912.

Porte, G. (Ed.) (2012). *Replication research in applied linguistics research.* Cambridge: Cambridge University Press.

Porte, G. (2015). *Commentary on replication research in applied linguistics research.* Discussant's colloquium paper at the *American Association of Applied Linguistics*. 21-24 March. Toronto, Canada.

Porte, G., & Richards, K. (2012). Replication in second language writing research. *Journal of Second Language Writing,* 21 (3), 284-293.

Piwowar H.A., Vision T.J. (2013). Data reuse and the open data citation advantage. *PeerJ* 1:e175 https://dx.doi.org/10.7717/peerj.175

Research Councils UK (2013). *RCUK policy on open access and supporting guidance.* Accessed 25 June 2015 from http://www.rcuk.ac.uk

Santos, T. (1989). Replication in applied linguistics research. *TESOL Quarterly, 23* (4), 699-702.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13 (2), 90-100.

Spinner, P. (2011). Review article: Second language acquisition of Bantu languages: A (mostly) untapped research opportunity. *Second Language Research,* 27, 418-430.

Trofimovich, P. (2015). Editorial: Rewarding Open Science. *Language Learning,* 65, 3.

VanPatten, B. (2015). Replication research in instructed SLA. Paper presented at the *American Association of Applied Linguistics*. 21-24 March. Toronto, Canada.

Vincent N. & Wickham C. (Eds.) (2013). *Debating Open Access*. London: The British Academy.

Wagner, A. B. (2010). Open access citation advantage: An annotated bibliography. *Issues in Science, Technology & Librarianship.* 60: Winter. Retrieved 25 June 2015 from http://www.istl.org/10-winter/article2.html

Wicherts, J., Bakker, M., Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results *PLoS* One , 6, 11, e26828.