Applied Linguistics 2023: 44/4: 718–746
https://doi.org/10.1093/applin/amad025 Advance Access published on 17 May 2023

Measuring the impact of academic literacy interventions: Refining an evaluation design through self-reflection and feedback

ILSE FOUCHÉ

Division of Languages, Literacies and Literatures, University of the Witwatersrand, Johannesburg, South Africa
*E-mail: ilse.fouche@wits.ac.za

This article, located in the discipline of academic literacy studies, draws upon the fields of critical realism, design research, and evaluation studies. It reports on the validation of a flexible evaluation design for assessing the impact of academic literacy interventions. The design was validated in two ways. Firstly, through a process of critical reflection, the researcher considers her own experience with applying the evaluation design to an academic literacy course; the weaknesses and limitations that emerged from this implementation are considered. Secondly, academic literacy specialists responsible for a wide variety of interventions in South Africa were consulted by means of a questionnaire containing both quantitative and qualitative questions. The purpose of this questionnaire was to determine to which extent the evaluation design could be applied to a variety of academic literacy interventions in various contexts. Recommendations regarding the refinement of the evaluation design are made, and a revised evaluation design is put forward.

INTRODUCTION

The goal of research should be to make a dual contribution to both theory and practice; this includes helping to solve the practical problems encountered by practitioners (Cole *et al.* 2005; Liem 2011; Alfaro-Tanco *et al.* 2021). Within the South African higher education context, there has been a proliferation of academic literacy (AL) interventions over the past decades; however, not enough research exists on determining the impact of these interventions (Fouché *et al.* 2016). One reason for this might be that doing so comprehensively is a difficult undertaking, and that few guidelines exist to help practitioners and researchers to do so. In an attempt to address this practical problem, and as part of a larger design research project (Buchanan 2001), Fouché *et al.* (2016) suggested a flexible evaluation design (see Figure 1, further discussed under 'A summary of the originally proposed evaluation design') which could be used by practitioners in the field to determine the impact of their respective AL interventions; no comprehensive tool then

© The Author(s) (2023). Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

Figure 1: Originally proposed evaluation design for academic literacy interventions.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

existed which practitioners could use when embarking on such an evaluation process. This evaluation design was subsequently implemented, as reported on in Fouché *et al.* (2017) and Fouché (2017).

However, as is argued in this article, it is not enough to generate an instrument, even if it is embedded in previous research, as was the case with the instrument under discussion. Part of responsible research is also to test and validate research artefacts. This article aims to contribute to the field of AL development in higher education by firstly critically reflecting on the previous implementation of an evaluation design which could be used to evaluate various AL interventions, and secondly improving the content and face validity of this instrument by obtaining feedback from a range of academic literacy practitioners. It does so to indicate how reflection and external feedback can contribute to an improved design artefact. To this end, this article uses the framework of *design research*, which centres around a viable artefact which has been created in the form of a model, method, or construct; once such an artefact has been developed, it is important that the design artefact's utility, quality and efficacy be demonstrated by rigorously evaluating it (Buchanan 2001; Cole *et al.* 2005; Hevner and Chatterjee 2010; Mandviwalla 2015; Meyers *et al.* 2018). The goal of this interdisciplinary article, drawing on the diverse fields of design research, critical realism, evaluation studies and AL, is to do just this, with the aim of ultimately presenting a revised flexible evaluation design that could be used across a range of AL contexts.

The research design process of the larger study followed a generate-test cycle which, as will be discussed further in the 'Theoretical framework' section, moves through the phases of the creation of the artefact, providing a chain of evidence, and evaluating the process (Mandviwalla 2015). In the creation phase, a variety of tools were proposed from which researchers could select a combination which was relevant to their particular AL evaluation contexts. As part of the second phase, providing a chain of evidence, I implemented the evaluation design by assessing the impact of a first-year AL course at a South African university (Fouché 2017; Fouché *et al.* 2017). During these two phases, the instrument's construct validity was determined. See the section, 'A summary of the originally proposed evaluation design', for a discussion of these two phases. The third phase, on which the current article reports, involves the evaluation of the instrument, which itself has two parts, namely self-reflection, as well as improving the internal and face validity through external feedback.

The two aspects of evaluation are elaborated on in the next section, followed by a discussion the theoretical framework covering critical realism and design theory. Thereafter, a brief account of the original design's background and development is given. This is followed by the evaluation phase of the generate-test cycle, which consists of (i) a reflection on the challenges that were experienced in implementing this design, and (ii) feedback from South African AL specialists. Finally, a revised evaluation design is proposed.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

SELF-REFLECTION, CONTENT AND FACE VALIDITY

The evaluation design under discussion is itself evaluated using two methods. Firstly, I report on my own reflection around the challenges encountered in the previous implementation of the design (thus its use in practice), and make suggestions regarding future implementation. Reflecting on a research process is a pivotal part of responsible scholarship. It is crucial to understand what worked, what did not work, and what the reasons therefor might be. Dewey's (1933) definition of reflection is still drawn on today. According to Dewey (1933: 6), '[r]eflection is an active, persistent and careful consideration of any belief or supposed form of knowledge in light of the grounds supporting it and future conclusions to which it tends'. Yost *et al.* (2000) add that reflection entails a belief (or disbelief) in something based on some sort of evidence. For example, when assessing the impact of a phenomenon, it is important to reflect upon the measures and methods used to assess such impact to ascertain whether these resulted in reliable and valid evidence, or whether they could be altered to obtain richer, more valid, and more reliable information in future; thus, to get closer to a true reflection of what happens in the Actual domain from within a critical realist ontology, as discussed in the next section. Only by doing this can scholarship in the field be improved upon. In my discussion I reflect on challenges associated with five main areas: the insider/outsider dilemma, external collection of data, using a generic AL test for pre- and post-testing, using an extended essay, and using a questionnaire.

The second method of evaluation involved determining the instrument's content and face validity by sending it to AL specialists across South Africa to establish whether the evaluation design was appropriate and adequate for their respective contexts, and how they would adapt it to be more appropriate. Responses were received from practitioners at 14 of South Africa's 26 public universities. I discuss quantitative and qualitative findings generated from the analysis of their responses.

I draw on these evaluation methods to rigorously evaluate the utility, quality, and efficacy of the instrument through a critical realist lens (cf. Cole *et al.* 2005; Hevner and Chatterjee 2010; Mandviwalla 2015; Meyers *et al.* 2018). My purpose is to indicate how reflection and external feedback, as part of the generate-test cycle, can contribute to an improved artefact with which to describe results observed in the Empirical domain—in the context of this study, an improved evaluation design within the framework of design research (Buchanan 2001). I end by presenting an adapted evaluation design which is strengthened for use by practitioners who wish to evaluate the impact of their own AL interventions, but who are unsure as to how this should be embarked upon in a responsible manner. The review of this specific evaluation design should be considered a case study to indicate the importance of reviewing an artefact as part of a cyclical process of continuous improvement. The revised evaluation design will be

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

of use to practitioners within the fields of academic writing and broader AL interventions.

THEORETICAL FRAMEWORK: DESIGN THEORY AND ITS ONTOLOGICAL BASE IN CRITICAL REALISM

The assumption that a designed artefact, such as the evaluation instrument under discussion, is able to describe (at least partially) an AL intervention's impact points to the ontology underlying this study, namely critical realism. In the context of this study, critical realism is considered useful for two reasons. Firstly, as Corson (1997: 168) aptly illustrates, critical realism is a suitable guiding philosophy for applied linguistics, in that the latter 'goes beyond the ideal concerns of linguistics itself. It steps resolutely into the ontological minefield that is the *real* world of human social interaction' (emphasis added). What is more, he argues that critical realism 'gives ontological status to human sign systems themselves' (Corson 1997: 183); the evaluation instruments that are at the core of this article are prime examples of such human sign systems. Secondly, the reflective capacity of critical realism (Dison *et al.* 2022) makes it particularly suitable for this study.

Critical realism posits that reality can be understood by considering three domains, namely the Empirical, the Actual, and the Real (Bhaskar 2008). The Real domain exists outside of human influence, but has causal powers which result in events in the Actual domain, where events occur that can be experienced. We attempt to describe these events in the Empirical domain—the observable world, or the world which we sense and can understand (cf. Corson 1997). Our description of these events can never be complete, as we can never fully access the Real domain. However, it is our responsibility as researchers to, inside the Empirical domain, come as close as we are able to in our attempts to describe or measure the events that happen in the Actual domain. Reflection and peer evaluation are two mechanisms that enable us to do this responsibly. In the generate-test cycle (originally used in the field of information technology in the 1960s, as seen in the work of Simon 1996), this reflection and evaluation would form part of the 'test' part of the cycle. From a critical realist perspective, reflection and evaluation are essential in proposing an instrument of transitive knowledge (Bhaskar 2008) that measures as closely as possible what happens in the Actual domain.

As the outcome of a generate-text cycle within the framework of design research (Buchanan 2001; Mandviwalla 2015), a revised evaluation design is proposed. This design research process can be aptly summarized in Figure 2, as proposed by Buchanan (2001), through the interplay between manufacturer (by means of design based on theory as well as self-reflection), product (the evaluation design artefact itself), and community of use (a wide range of AL practitioners). As Buchanan (2001: 14) states, the 'product then is a negotiation of the intent of the designer and manufacturer and the expectations of communities of use'.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

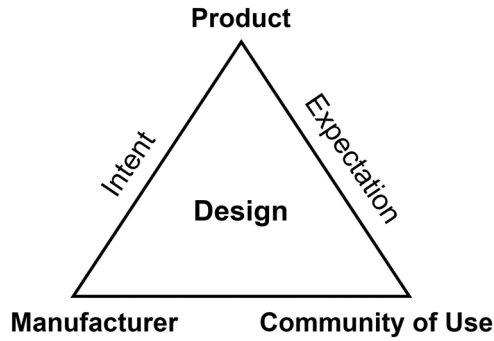

Figure 2: The process of design (Buchanan 2001: 15).

BACKGROUND TO THE DEVELOPMENT OF THE EVALUATION DESIGN

This study was conducted in South Africa, where the term 'academic literacy interventions' is generally used to refer to the vast array of interventions which are aimed at improving university students' ability to write, read, and otherwise engage with academic discourse. This study uses Van Dyk and Van de Poel's (2013: 56) view of AL as 'being able to use, manipulate, and control language and cognitive abilities for specific purposes and in specific contexts'. This generic definition acts as an umbrella term for the wide range of AL interventions that were identified by participants in this study. Of course, each one of those might choose a narrower definition of AL depending on their own theoretical orientations, objectives, and contexts. However, this definition does seem to encompass a broad and often contested construct without excluding any ideological approaches to AL, such as the very influential New Literacies Studies movement or more skills-based approaches. It should be noted that the construct of AL has come to be seen as much more than language abilities. Rather, many if not most researchers in the field would describe AL as multiple (thus, as academic literacies), embedded within specific disciplines (thus requiring an understanding of these disciplines' epistemologies and conventions), being laden with power and embedded in socio-cultural contexts, and linked to identity construction (Gee 2008; Wingate 2015). What is more, the concept of 'developing' AL abilities (as was alluded to in the introductory section of this article) is contentious, and often compared to 'learning' these abilities, with the former referring to the long-lasting integration of these abilities into the student's repertoire, and the latter referring to a more temporary acquisition that can be reversable (cf. Granott 1998). As will be argued later in this article, the distinction between development and learning needs to be taken into consideration when assessing impact.

In part due to historical inequalities, and in part due to the global trend of massification and the often concomitant decline in the preparedness of the

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

bulk of university students to write at the level required for university studies, AL is a major area of focus in South African higher education (cf. Wingate 2015). Most South African public universities currently have some form of AL intervention in place, be it credit-bearing courses or other interventions such as writing centres (see, amongst others, Pienaar 2005; Ngwenya 2010; Van Dyk *et al.* 2011; Winberg *et al.* 2013; Van Wyk 2014; Bharuthram and Clarence 2015; Boakye and Mai 2016; Dison and Mendelowitz 2017; Fouché *et al.* 2017; Carstens and Rambiritich 2021). Despite the abundance of AL interventions, very few studies have attempted to evaluate the impact of these interventions. This might be because of the difficulty of capturing in the Empirical domain what happens in the Actual domain, specifically with regard to interventions which operate within larger systems with numerous aspects that influence their impact. Indeed, Scott (2005) indicates that it is exactly because of the fallibility of human understanding of how the world truly works that researchers must critically and continuously reflect on phenomena and their own ways of describing these.

Determining the impact of an intervention falls within the sphere of evaluation. Babbie (2021: 358) defines evaluation research as the 'process of determining whether a social intervention has produced the intended result'. De Vos *et al.* (2011) argue that, due to an increased focus on accountability, stakeholders want to see proof that interventions work, of how they work, and of how they can be improved. Lynch (2003), whose research focuses on the evaluation of language programmes, states that programme evaluation draws on the data gathered from language assessments so as to reflect on the intervention in question, make decisions and take appropriate actions based on the feedback received. Bachman and Palmer (2010) agree that in the process of evaluation, language assessments are used primarily to inform the decisions and value judgements that are made about a specific programme, intervention, or construct.

As mentioned in the introduction, it was to address this gap in the literature on assessing the impact of AL interventions, as part of a generate-test cycle within a design research framework (Buchanan 2001; Mandviwalla 2015), that I originally developed the evaluation design (Fouché *et al.* 2016) illustrated in Figure 1, and which I summarize below.

A SUMMARY OF THE ORIGINALLY PROPOSED EVALUATION DESIGN

Lynch (1996) states that an evaluation design can be seen as a methodological strategy for evaluating programmes. Bamberger *et al.* (2012) add that an evaluation design is a plan which includes both sources and methods of data collection as well as methods of analysis. Judd and Keith (2017) argue that outcome data alone (e.g. data which show an improvement in students' AL levels) may not be sufficient to draw conclusions that the intervention

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

caused the outcome, but that causal inferences become more justified if data are collected in a framework with other sets of data—thus, if data are triangulated.

Fouché *et al.* (2016) proposed that, in order to comprehensively assess the impact of an AL intervention, two aspects should be explored by means of an evaluation design. These two aspects are included in the definition of impact in the context of AL interventions put forward by this study, namely that impact is '(i) the observable improvement in AL abilities between the onset and the completion of an AL intervention, and (ii) the extent to which these abilities are necessary and applied in students' content subjects'. The second of these two aspects is particularly important in the field of AL, and sets it apart from general language evaluation, as well as evaluation studies in general. AL interventions are always aimed at empowering students by improving their ability 'to use, manipulate, and control language and cognitive abilities for specific purposes and in specific contexts' (Van Dyk and Van de Poel 2013: 56)—if students are ultimately not able to transfer the improved levels of AL displayed in the AL intervention to relevant contexts, the intervention cannot be considered successful. An additional dimention to this is considering whether students truly develop the relevant AL abilties, and are able to draw on these effectively not only in other contexts, but also in later years.

Logistical and ideological factors influence the type of AL intervention which programmes, departments, faculties, or universities are able to put in place. Sometimes, generic interventions are necessary due to structural limitations such as timetabling issues and a diversity of students who participate in the intervention. For the purposes of this article, generic interventions are differentiated from a 'study skills' approach in that the latter, coined by Lea and Street (1998), operates from a deficit-based approach which attempts to correct students' writing, usually by teaching writing as a formula (Clarence and McKenna 2017). The term 'generic academic literacy interventions', in this article, rather indicates that the intervention services students from multiple disciplines; such interventions might draw on a study-skills approach, but they might also be positioned within other epistomologies within the field of academic literacies. Proponants of a New Literacies Studies approach, however, will make a strong case for discipline or even subject-specific interventions, as this movement has compellingly argued for the necessity of acquiring a D/discourse (cf. Gee 2008; Wingate 2015) within discourse communities. Some contexts allow for a collaborative approach to strenghtening students' AL (Wingate 2015), by having AL experts and content experts work closely together, sometimes even teaching in the same classroom. Clearly, the environments and realities of institutions and departments differ, and as Judd and Keith (2017) point out, these different environments will naturally impact on outcomes. Mhlongo (2014) echoes this sentiment and warns that the context of AL interventions must be kept in mind when these are evaluated, due to the unique challenges faced by various higher education institutions.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

The proposed evaluation design was therefore flexible in that evaluators could choose a combination of tools from a variety of evaluation instruments to suit their specific contexts. Thus, different combinations of evaluation instruments could be used to best describe, in the Empirical domain, the realities of the Actual domain for each intervention (cf. Bhaskar 2008). These evaluation instruments included generic or subject-specific AL tests, generic or subject-specific extended writing assignments (evaluated either by means of a rubric or quantitative measures), student and lecturer questionnaires, qualitative feedback from primary stakeholders, content analyses of study material, and correlating AL achievements with other variables. Using a combination of instruments, some of which were recommended and some of which were considered optional, would ensure that data are triangulated by both method and source, and would also contribute to the validity of findings (Jick 1979; Lynch 2003; Judd and Keith 2017). In the original evaluation design, it was proposed that, where it was possible to use control groups, at least two instruments be used, and in cases where control groups were not possible, at least three instruments be used. Regardless of which combination of instruments was decided upon, it was recommended that at least one measured whether students' AL abilities had improved, and at least one measured whether these abilities were required in, and/or transferred to, students' other subjects. By using several instruments that measure impact from different perspectives, triangulation by both method and source becomes possible. Different combinations of these instruments were proposed for four broad types of course, namely generic AL interventions, subject-specific AL interventions, collaborative AL interventions, and limited-purpose AL interventions (e.g. writing centres or reading programmes) (cf. Van De Poel and Van Dyk 2015). The originally proposed instruments are indicated in Figure 1.

It should be noted at this stage that this study does not make a case for one ideological approach over another. Its aim was to propose an evaluation design that could be used across a range of interventions and ideologies, with an arsenal of instruments from which a researcher could select relevant tools to assess the impact of specific interventions, with guidelines of doing so comprehensively. However, some of the instruments proposed in the evaluation design might seem to favour a skills-based approach. If this is the case, that might be because more literature (which formed the foundation of the initially proposed evaluation design) exists on evaluative approaches to AL interventions from that theoretical orientation. Lillis (2003: 192), for example, states of the New Literacies Studies movement that '[w]hilst powerful as an oppositional frame, that is as a critique of current conceptualisations and practices surrounding student writing, academic literacies has yet to be developed as a design frame (…) which can actively contribute to student writing pedagogy as both theory and practice'. Not much has changed in the intervening two decades in terms of practical suggestions for evaluative approaches within this framework.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

This section has described the 'creation' phase of the evaluation design. The next section starts by briefly summarizing the context of the second phase of the generate-test cycle. The focus of the following section is on the first part of the third phase of this cycle (and the focus of this article), namely evaluating the process (cf. Mandviwalla 2015) by critically reflecting on the challenges experienced in implementing the evaluation design.

IMPLEMENTATION OF THE ORIGINAL EVALUATION DESIGN

Before the challenges experienced are expanded upon, the context of the implementation of the evaluation design, as discussed in detail in Fouché *et al.* (2017) and Fouché (2017), is summarized here. The evaluation design artefact was implemented to evaluate the effectiveness of a generic, two-semester long AL course which was aimed at first-year students from a wide variety of faculties at a South African public university. The AL intervention required students to attend two 50-min classes per week, and addressed the following outcomes:

1. Identifying word meaning from context;
2. Paraphrasing text;
3. Making effective notes from presentations and reworking these notes to paragraphs and mind maps;
4. Including references in a text;
5. Understanding academic genres and identifying and finding reliable academic sources;
6. Explaining the concepts of active reading, skimming, and scanning;
7. Using skimming and scanning to obtain information from texts;
8. Identifying the qualities of, and being able to write good introductions and conclusions;
9. Creating a table of contents, and using it to plan and structure text;
10. Writing paragraphs with clear topic sentences, one main idea, and applicable support;
11. Identifying action words and content words in examination questions and assignments, and planning well-structured responses to examination questions;
12. Identifying reasons for using the passive voice;
13. Identifying inaccurate information;
14. Writing correct sentences;
15. Calculating basic percentages;
16. Explaining and being able to identify visual manipulation;
17. Referring correctly to different parts of graphs and tables;
18. Identifying reasons for using graphic information, analysing graphics, and discussing graphics appropriately;
19. Being aware of the structure of a seminar, being able to ask effective questions, and being able to answer questions effectively; and
20. Distinguishing between open, closed, and hypothetical questions.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

Findings from this evaluation indicated that the AL course effectively addressed a wide range of AL abilities, though important areas such as students' vocabulary and ability to avoid plagiarism showed little improvement. Furthermore, large effect sizes were only evident after a full year's intervention, indicating that shorter interventions are unlikely to be sufficient in improving students AL levels. The course which was selected for evaluation was not presented at my university, and was also in another province in South Africa. I was therefore not on site during this process, and had to rely on the course coordinator to gather the necessary data.

After consultation with course developers, the following instruments were selected for assessing the impact of this course: a generic AL test, a generic extended writing assignment (assessed by means of both a rubric and quantitative measures), a student questionnaire, and correlating AL achievements with other variables. As the course was generic in nature (due to the wide variety of students it serviced), subject-specific instruments were not viable. For the same reason, content analysis of study material as well as lecturer questionnaires were seen to be impractical—the vast number of stakeholders and the large number of content-subjects (97 in total) made these instruments very difficult to implement effectively.

CHALLENGES EXPERIENCED IN IMPLEMENTING THE DESIGN: CAVEATS AND REFLECTIONS

'Through praxis, critical consciousness develops, leading to further action' (Baum *et al.* 2006: 856). Praxis here is differentiated from the practice of teaching, in that praxis is the integration of practice and theory, a 'deliberative, responsible, human-moral action' which involves 'wise judgement', in the Aristotelian sense (Connor 2004: 56). The suggested evaluation design was accordingly based on theory and best practice grounded in the literature. Yet, as part of a generate-text cycle (Mandviwalla 2015) and as pointed out by Baum *et al.* (2006), theory must be put to the test in praxis. The 'critical consciousness' and reflection that result from this process are crucial for further improvement (cf. Yost *et al.* 2000), in the case of the current study, of an evaluation design as artefact. At the same time, it is important to acknowledge some caveats which impacted the study, but which, depending on context, might be difficult to alter in subsequent studies of this nature. These include possible restrictions on whether insiders or outsiders collect data, the realities surrounding the use of extended writing assignments and the challenges with online questionnaire response rates.

My reflection on the challenges encountered in implementing the evaluation design considers five main points. These should be kept in mind by future researchers wishing to use the same instruments.

THE INSIDER/OUTSIDER DILEMMA

A preliminary point that should be considered before undertaking an evaluation of an intervention is that of insider and outsider roles with regard to

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

evaluation studies (Hawkey 2006). In evaluating the impact of this course, I was what would be defined as an outsider. I do not work at this specific university, nor did I have a stake in how students perform in the evaluations. The fact that I was not on site during the evaluation period meant that I relied on the assistance of course coordinators to gather data on my behalf. Although I did not have a stake in how students performed in the evaluations, I did have a stake in how well the instruments were implemented, and in the evaluation process. Being an outsider created several challenges in this regard. An insider (and specifically a course coordinator) would have had more flexibility in structuring assessments in terms of timing and frequency to enable optimal evaluation of the intervention. Furthermore, being in closer contact with the students and course lecturers would likely have resulted in better relationship and trust building (Kerstetter 2012; Muhammad *et al.* 2015), which in turn might have resulted in higher participation rates and consequently improved data gathering (Gasman and Payton-Stewart, 2006; Chawla-Duggan 2007). I would therefore recommend that where possible, future studies assessing the impact of AL interventions either have insiders conducting the evaluation, or have outsiders who can be on site and closely involved during the evaluation period.

EXTERNAL COLLECTION OF DATA

Closely linked to the insider/outsider dilemma was the fact that in this research study, data was collected by people other than the researcher. This might have led to ethical dilemmas in some cases. For example, had the staff members who collected the data had a bigger stake in the outcome of the evaluation design (e.g. the risk of losing their jobs), they might have been tempted to interfere with the data. In this study, that was not the case. Furthermore, the researcher was in close contact with the course lecturers, and was copied in on all communication with students. The problem encountered in this case was rather that course lecturers were perhaps not invested enough in the research study, and therefore had no great stake in motivating all students to participate. In either case, it would have been preferable for me as researcher to collect the data myself, and this should be kept in mind for future similar studies.

THE USE OF AN AL TEST AS PRE- AND POST-TEST

A further challenge encountered involved the use of a generic AL test as a pre- and a post-test. The Test of Academic Literacy Levels (TALL) was used as a pre-test. This test has been thoroughly validated and shown to be reliable (Van Rooy and Coetzee-Van Rooy 2015), and was thus a judicious choice as a pre-test. The AL course's two internal examinations were used as post-tests; one was written in May, the other in November.

Due to practical constraints, in particular time constraints, it was not possible to have students write both the TALL and the subject's examinations at the

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

end of each semester. As new examinations are set each year, it is not possible to subject them to the same process of validation as is the case with the TALL. However, as indicated in Fouché *et al.* (2017), the tests are based on the same construct and include the same subsections. In addition, Spearman's rho was used to draw non-parametric correlations between the tests, and moderate to strong correlations between the tests indicated that the tests in their entirety could be considered equivalent through statistical measures. What is more, students' improvement between pre- and post-tests was shown to be statistically significant. It is important to note, though, that internal reliability rates of some of the examinations' sub-sections were not high enough to make it possible to compare the sub-sections of the respective tests so as to ascertain in which sections students had improved the most.

Previous studies have indicated that it is highly unlikely that students AL levels will improve meaningfully without explicit instruction (see, e.g. Thompson 1990; Farnill and Hayes 1996; Rosenthal 1996; Holder *et al.* 1999; De Graaff and Housen 2009). Though it cannot be proven without a doubt, it is reasonable to assume that the improvement in AL levels can be attributed to the intervention. Thus, though it could be safely concluded that students' AL levels as a whole improved between the TALL and the two post-tests, and that this improvement could likely be attributed to the intervention, it was not possible to determine which specific AL abilities improved. This makes it difficult to establish how the curriculum could best be adapted in future so as to optimally improve students' AL abilities. A recommendation for future research would therefore be to ensure that the same test be used as both pre- and post-test. Where this is not possible, tests should ideally be piloted to ensure that they (and their various sub-sections) can be considered theoretically and practically equivalent before any valid conclusions can be drawn based on results from these tests.

THE USE OF AN EXTENDED WRITING ASSIGNMENT

Another set of challenges was encountered in assessing generic extended writing assignments written by students before, in the middle of, and at the end of the intervention. These challenges expanded upon in Fouché (2017) included obstacles in collecting assignments from students in the original randomly selected sample, high student numbers combined with limited resources and curriculum credit constraints influencing the type and number of assignments that could be given to students, and very few writing assignments being prescribed to first-year students in other content subjects (likely due to high student volumes). Even with the advantage of hindsight, these challenges would have been difficult to overcome. However, some additional measures might help mitigate similar challenges for future researchers. These include ensuring that effective communication mechanisms between lecturers and students are in place, as well as setting up collaborations between lecturers from students' other subjects well in advance, to co-design collaborative writing assignments.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

A further challenge regarding the written assignments concerned assessing these assignments by means of quantitative measures. Specifically assessing accuracy scores is extremely work-intensive, and it is not practically possible to do this effectively unless significant resources (time as well as money for several well-qualified markers) are available. In this study, it was only possible to use this instrument with a sample of 50 randomly selected scripts from the AL intervention that was evaluated in the Fouché (2017) study, and even that took several weeks to complete. It should thus be kept in mind by future researchers that this instrument cannot be used effectively with large student populations without accompanying resources. Further, this instrument is better suited to a study-skills approach, which has fallen out of favour by most current practitioners in the field. Though it is still included as an option in the final evaluation design, prospective course evaluators should carefully consider the value of this instrument before including it in their own evaluation designs.

THE USE OF QUESTIONNAIRES

A final challenge that was encountered was with regard to the student questionnaire. The response rate to the first round of questionnaires at the beginning of the course was extremely low—a result consistent with existing research on online questionnaires (Saleh and Bista 2017). Due to the problems experienced in the first semester, hard-copy questionnaires were handed out in class at the end of the second semester. This yielded better results, with 84 of 976 students completing the questionnaires. This number was still fairly low, possibly due to relatively few students attending classes in the last week of the semester, yet the sample was large enough to draw deductions from (at a 90 per cent confidence level with an 8.6 per cent sampling error level—see Dillman 2007). It is thus recommended that provision is made to distribute questionnaires in class—late enough in the semester for the curriculum to have been completed, but at a point where most students still attend class.

Many of the challenges highlight the need for meticulous planning before a large-scale evaluation of the impact of an intervention is undertaken. Even if this is done (as was the case in this study), unforeseen obstacles are likely to arise which will necessitate the implementation of alternative plans. By continuously reflecting on new challenges that arise, the accompanying critical consciousness that will develop (cf. Baum *et al.* 2006) will enable the researcher to keep refining theory, which can in turn be positively implemented in praxis. The following section reports on the feedback received from AL specialists from across South Africa on the usefulness and relevance of the proposed evaluation design.

FEEDBACK FROM AL SPECIALISTS

Cattani *et al.* (2014), in discussing the evaluation of artefacts in cultural fields as diverse as art or science, state that the audiences relevant to a specific field

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

must offer subjective evaluations thereof for the artefact to be considered culturally legitimate and gain symbolic capital. It was therefore important to obtain subjective feedback from a broad range of AL practitioners as part of the generate-test cycle (Mandviwalla 2015) for the evaluation design artefact which is the focus of the current study to gain such legitimacy within a fairly diverse field, both in terms of underlying ideologies as well as types of interventions offered. To this purpose, a questionnaire was designed to determine whether the proposed evaluation design met the needs of AL specialists across a broad spectrum of universities and interventions. Responses were obtained from 14 of the 26 universities in South Africa. In several cases, where universities ran multiple AL interventions, two or three AL specialists from a university responded. In total, 23 responses were received, which accounts for a response rate of approximately 50 per cent. These responses cover a broad range of AL interventions. Seven specialists who responded were responsible for generic undergraduate AL courses, 14 were responsible for subject-specific AL courses, two were responsible for generic postgraduate AL interventions, three were responsible for subject-specific postgraduate AL interventions, four were responsible for collaborative AL interventions, eight were responsible for writing centres, and one was responsible for a reading laboratory. The number of interventions indicated above (39 in total) is greater than the total number of responses (23) as several specialists are responsible for more than one type of intervention. Though a larger response rate would have been preferable, and results in this section therefore need to be treated cautiously, the responses still provide valuable insight into the usefulness of the evaluation design.

QUANTITATIVE FEEDBACK FROM AL SPECIALISTS

The AL specialists were first asked whether the various proposed instruments were 'very applicable', 'applicable to a limited extent', or 'not applicable' to their respective contexts. As can be seen from Table 1, the majority of respondents felt that each of the proposed instruments were either 'very applicable' or 'applicable to a limited extent'.

Further, an attempt was made to establish whether certain instruments were more applicable to certain types of AL interventions. However, a limitation of the questionnaire was that it did not allow AL specialists who are responsible for multiple interventions to distinguish between these interventions when reporting on the applicability of instruments. To account for this limitation, only the responses of AL specialists who are responsible for only one intervention (n = 16) are used in Table 2. Only four types of interventions are represented by this reduced sample, namely generic undergraduate AL courses (abbreviated Gen UG AL), subject-specific undergraduate AL courses (abbreviated SS UG AL), collaborative AL interventions (abbreviated Collab AL), and writing centres (abbreviated WC).

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

Table 1 : Academic literacy specialists' perceptions of usefulness of instruments

Instruments	Very applicable	Applicable to a limited extent	Not applicable
Generic academic literacy test	8	8	7
Subject-specific academic literacy test	14	6	3
Generic extended writing assignment (rubric)	4	12	7
Subject-specific extended writing assignment (rubric)	17	4	2
Quantitatively assessing writing assignment	9	9	5
Student questionnaire	14	8	1
Lecturer questionnaire	9	12	2
Content analysis of study material	13	8	2
Correlating academic literacy results with other variables	11	7	5
Qualitative feedback from primary stakeholders	21	2	0

Although only approximately half of the original responses could be used for the more detailed information presented in Table 2, the same trends as in Table 1 are evident. Although all instruments are either very applicable or applicable to a limited extent to some AL interventions (meaning that none can be discarded in the revised evaluation design), there does seem to be a strong preference for subject-specific instruments (even among generic interventions—it is possible that even these interventions are integrating subject-specific elements into their curricula). Furthermore, there is an overwhelming preference for qualitative feedback from primary stakeholders, indicating a need for thorough, descriptive, and explanatory instruments which can be used to understand the impact of AL interventions. Such instruments might need to be separately developed for each specific evaluation context. Future research could however consider whether credible, transferable, and dependable templates aimed at various primary stakeholders could be developed.

QUALITATIVE FEEDBACK FROM AL SPECIALISTS

Though the quantitative feedback indicates that all instruments in the proposed evaluation design would be either 'very applicable' or 'applicable to a limited extent' for the majority of respondents (see Table 1), qualitative feedback as part of the generate-test cycle used within the framework of design research (Buchanan 2001; Mandviwalla 2015) in this study was essential in

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

Table 2: Academic literacy specialists' perceptions of usefulness of instruments, categorized into types of interventions

Instruments	Type of academic literacy intervention	Total number of responses	Number of academic literacy specialists who found the instrument…		
			Very applicable	Applicable to a limited extent	Not applicable
Generic academic literacy test	Gen UG AL	**4**	3	0	1
	SS UG AL	**8**	2	5	1
	Collab AL	**2**	0	0	2
	WC	**2**	0	1	1
Subject-specific academic literacy test	Gen UG AL	**4**	1	2	1
	SS UG AL	**8**	8	0	0
	Collab AL	**2**	0	1	1
	WC	**2**	0	1	1
Generic extended writing assignment (rubric)	Gen UG AL	**4**	2	2	0
	SS UG AL	**8**	0	7	1
	Collab AL	**2**	0	0	2
	WC	**2**	0	1	1
Subject-specific extended writing assignment (rubric)	Gen UG AL	**4**	2	2	0
	SS UG AL	**8**	8	0	0
	Collab AL	**2**	1	1	0
	WC	**2**	1	0	1
Quantitatively assessing writing assignment	Gen UG AL	**4**	2	2	0
	SS UG AL	**8**	3	4	1
	Collab AL	**2**	0	1	1
	WC	**2**	1	0	1
Student questionnaire	Gen UG AL	**4**	2	2	0
	SS UG AL	**8**	6	2	0
	Collab AL	**2**	0	2	0
	WC	**2**	2	0	0
Lecturer questionnaire	Gen UG AL	**4**	1	3	0
	SS UG AL	**8**	6	1	1
	Collab AL	**2**	0	2	0
	WC	**2**	0	2	0
Content analysis of study material	Gen UG AL	**4**	4	0	0
	SS UG AL	**8**	6	2	0
	Collab AL	**2**	0	1	1
	WC	**2**	0	2	0
Correlating academic literacy results with other variables	Gen UG AL	**4**	3	1	0
	SS UG AL	**8**	4	3	1
	Collab AL	**2**	0	1	1
	WC	**2**	0	0	2

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

Table 2. Continued

Instruments	Type of academic literacy intervention	Total number of responses	Number of academic literacy specialists who found the instrument…		
			Very applicable	Applicable to a limited extent	Not applicable
Qualitative feedback from primary stakeholders	Gen UG AL	**4**	4	0	0
	SS UG AL	**8**	8	0	0
	Collab AL	**2**	1	1	0
	WC	**2**	2	0	0

understanding how some instruments could be refined, and whether there was a need for additional instruments. This was done by asking AL specialists three questions which are discussed in the remainder of this section. Themes identified through a process of content analysis are discussed below.

Improving instruments

The first question posed to AL specialists was 'Are there any ways in which you believe any of the proposed instruments could be improved?'. Half (n = 10) of the participants who answered this question indicated that the instruments were 'suitable' and 'addressed the intended outcomes'. Several of the participants did, however, have suggestions as to how some of the instruments could be improved. The first comment was that the student questionnaire was too long, and that *this might serve as a deterrent to student response*. This might be an additional reason for so few electronic student questionnaires being returned after the first semester of the course. The original 48 questions were condensed to 33 questions in a revised questionnaire (Appendix A). To ensure that the student and lecturer questionnaires still correspond so as to facilitate triangulation, the latter (see Appendix B) was similarly revised. Another participant highlighted the fact that the lecturer questionnaire would *have to be explained to the lecturer, and close co-operation will be necessary*—this corresponds to Jacobs' (2005) findings that AL specialists can assist subject specialists by helping them make the tacit knowledge of the latter overt and explicit in their teaching practice. As is the case with the student questionnaires, this also indicates the importance of personal contact (i.e. facilitating the completion of the questionnaire), which again raises the insider/outsider dilemma. Yet, such personal contact is not always feasible due to the resources required. Trade-offs will thus have to be made, depending on the resources available and the evaluation context.

One participant commented on the appropriateness of the writing rubric for postgraduate students, indicating that more categories, for example genre and audience, might be needed. The writing rubric (Appendix C) has been adapted to reflect these suggested categories. This participant also indicated that the rubric

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

was *very limited for the types of genres that students have to produce*. This is a valid point—the rubric is aimed at relatively standard essay-type assignments, and not, for example, at genres such as laboratory reports or legal writing. However, developing appropriate rubrics for every type of academic writing genre lies outside the scope of the current study. The instrument provided is a guideline for what might be considered the most common type of academic writing, and it was proven useful in assessing subjects such as history, social anthropology, geography, physiology, and urban morphology (Carstens 2009). However, to assess specialized writing genres, alternative rubrics are likely to be needed.

Two participants proposed that the rubric be adapted to enable formative assessment in addition to summative assessment. One of these participants indicated that the draft method could be used in impact measurement—thus, that the impact of an intervention on students' writing be assessed through several drafts. One challenge with regard to this suggestion is that extensive resources (e.g. time and qualified markers) would be necessary for marking several drafts, especially in the South African context where classes are frequently large, and lecturers tend to have heavy teaching loads. Yet, there might be contexts where the required resources would be available, or where small class sizes allow for additional marking. In such contexts, this would seem to be a valuable addition to the evaluation design, and the writing rubric (Appendix C) was adapted by adding more details to facilitate formative assessment.

Another suggestion with regard to using writing assignments is that group assignments be used. This seems to be a sensible suggestion, considering how little individual writing seems to be done by students in their first year, especially where large student numbers are a reality. This would have to be thoroughly planned and managed though (for an example of best practices in this regard, see Michaelsen and Sweet 2011), so that both pre- and post-assignments are similar, and so that all students contribute equally to the end product. Furthermore, the group members for both pre- and post-assignments would ideally have to stay consistent so that various groups' assignments could be compared with each other. Without proper planning and management, the reliability of this instrument would suffer; appropriate checks and balances throughout the group work process would be essential.

Finally, one participant suggested that an assessment be included which assesses postgraduate students' *ability to identify arguments in the texts they read for research projects.* The TALL is an appropriate instrument for assessing undergraduate students' AL levels, and includes questions which assess students' ability to identify arguments in texts. However, as the proposed evaluation design should be adaptable for both undergraduate and postgraduate students, it would be sensible to propose an AL test aimed at postgraduate students, which includes abilities such as the one that this participant raised. The Test of Academic Literacy for Postgraduate Students (Butler 2009; Rambiritch 2013) is a widely used, validated test aimed specifically at postgraduate students, and is therefore suggested as an appropriate instrument to add to the current arsenal of instruments.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

Additional instruments

The second question asked was 'Are there any instruments that you would add to this evaluation design?'. Although the majority of participants felt that the range of tools was sufficient to assess the impact of their respective interventions, some valuable suggestions were provided for possible additional tools.

One participant indicated that they would have included listening and speaking abilities in the generic or discipline-specific tests. The participant specifically focussed on vocabulary acquisition: whether the student would be able to understand a word when hearing it in context, or whether the student would be able to use new words correctly when speaking, for example in a class conversation or a debate where the student does not have much time to think about the usage of the word. It is interesting to consider why the originally proposed evaluation design did not include assessment instruments for listening and speaking. If we consider the definition of AL proposed by Van Dyk and Van de Poel (2013), as described earlier in this article, all modes of communicating in an academic context, including speaking and listening, should be equally valued. The fact that I had not considered these in the original design indicates my own bias that academic communication should be centred around printed text (thus, reading and writing). This bias is not isolated. In fact, writing is frequently privileged in AL development because this remains the main mode through which students are assessed (Weigle 2002; Archer 2008; Van Dyk *et al.* 2009). Interventions which take a New Literacies Studies approach, in particular, tend to focus on writing (Lea 2004). Regardless of this reality, a comprehensive evaluation design must be exactly that: comprehensive. It was therefore necessary to consider how an assessment of speaking and listening could be interwoven into the evaluation design.

Though it would be possible to integrate such abilities in custom-made discipline-specific tests, it is not feasible to do this in pre-existing, widely used validated and reliable generic AL tests. Still, as some AL interventions do have outcomes that are assessed through the modes of listening and speaking, it would seem useful to add tools to the evaluation design which could be used to assess these abilities. A possible rubric for assessing oral presentations is attached as Appendix D. This rubric is taken from a course called Language and Study Skills which is presented at the University of Pretoria (Fouché and Immelman 2015), and is aimed at assessing formal oral presentations for first-year students. It might thus not be applicable to all contexts. Further research would be necessary to determine whether it could be used across a wider range of contexts. For AL interventions that have listening abilities as a major outcome, it would be valuable to consider the research done by Marais and Van Dyk (2010), who propose a listening test (the Academic Listening Test) aimed at first-year students which was shown to have construct, content, and face validity, in addition to being shown to be reliable. The revised evaluation design includes the possibility for assessing students' speaking and listening abilities.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

A further suggestion was that qualitative student data be used. Although the originally proposed student questionnaire did contain some open-ended questions, several other possibilities exist to collect rich and detailed information *so as to obtain a more holistic picture of development as well as the nature of students' challenges with the literacy practices of their discipline*. Firstly, the revised need-press questionnaire on AL abilities (Appendix A) was supplemented with an additional open-ended question so as to facilitate richer data collection. This question reads: 'How could the AL course be adapted to be of more value to you in your other subjects'. Further options which were suggested by participants include retrospective self-observation blogs as well as student interviews. To this, one could add other options such as focus group interviews. Although the originally proposed evaluation design did include *qualitative data from primary stakeholders*, the stakeholders focussed on were lecturers. In the revised evaluation design, a distinction is made between qualitative feedback from (AL and content-subject) lecturers and from students.

A final instrument that was suggested was benchmarking with other institutions. This can indeed be useful in finding out whether students at various institutions are at similar levels with regard to AL, whether institutions focus on similar outcomes in their AL interventions, and whether academic interventions at some institutions seem to be more effective than those at others. At the very least, such benchmarking could assist in identifying gaps and weaknesses (or strengths) in an institution's own AL intervention(s). At best, this might lead to collaboration and shared expertise, which would in turn benefit the interventions involved. As such, it is added as an additional tool to the proposed design.

Additional aspects to keep in mind

The third question posed to AL specialists was 'Are there any other aspects that you believe the proposed evaluation design should take into consideration, given the context of the AL intervention you are responsible for?'. Several constructive responses were received for this question.

Three participants commented on the duration of the AL intervention or the duration of the evaluation of its impact (cf. Bamberger *et al.* 2012). As one participant noted, instruments *would have to be longitudinal in nature, tracking the progress and development of AL skills over a period of time*. This comment is essential in our understanding of students learning AL abilities, as opposed to truly developing them (cf. Granott 1998). Where at all possible, students should be tracked longitudinally to determine whether they are still able to draw on the abilities acquired in the AL intervention in subsequent years.

Another participant pointed out that such an impact evaluation should not be *once off*—it is necessary to regularly evaluate the impact of AL interventions so as to ensure that they still fulfil their outcomes optimally. One participant commented on the fact that the *length of the intervention has an impact on the extent to which students can be said to have successfully learnt the literacies of their*

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

learning contexts. As has been confirmed by Van Dyk *et al.* (2011) and Fouché *et al.* (2017), the impact of AL interventions is often only noticeable after a longer intervention. The participant echoes this when stating that the *courses we teach are offered for a semester only due to timetabling issues, [and] resources in the form of teaching staff who are available to offer more extended interventions. Student feedback has indicated that they would have benefitted from a whole year rather than a semester*. The duration of the intervention must thus be kept in mind when the impact thereof is reported upon.

Other aspects that were raised by some participants were the background as well as the academic strength of the students. With regard to background, it would be important to consider students' home languages and their proficiency in these languages. In the South African context, only 9.6 per cent of the population speaks English at home (based on the most recently available 2011 census data from Lehohla 2012), while most learners are educated in English, especially at secondary level. As a result, these learners' reading and writing proficiency in their own home languages have often been neglected, which could further disadvantage them acquiring AL through the medium of English (Yamashita 2002; Ardington et al 2020). The needs of such students might differ from those of students who are fully proficient in another language. An argument to be made for a focus on reading and writing might be that the typical South African second language speaker of English might have had more exposure to the language in its oral form than typical English second language speakers, and that it is therefore necessary to focus on the language in its print-based mode, which is heavily drawn on in higher education. However, it falls outside of the scope of this article to evaluate which aspects of the AL construct should be focussed on AL conventions such as those in South African contexts. What is important is that the proposed evaluation design is able to address all the aspects of AL that course developers of various interventions deemed to be suitable for their own contexts.

Another participant stated that a *strong student can easily pick up a skill, whether the intervention is good or bad. It is the poorer student, however, whose improvement is proof of a good intervention*. This claim could be assessed by categorizing students into various quartiles based on pre-intervention data (e.g. Grade 12 or pre-test marks). These data could also indicate whether the intervention has the same impact on students at various levels, and whether it might be necessary to adapt interventions for students of different academic levels.

A further aspect that should be kept in mind is the planning stage of the evaluation. Lynch (1996) argues that when evaluating a programme, the first step would be identifying and consulting with relevant stakeholders. One participant echoes this notion: *I would first determine from the stakeholders what they will need, for example what does the ethics committee require? What does the disciplinary supervisor require? What about the funders?* Answering these questions would be pivotal in selecting suitable evaluation instruments. A related comment made by a participant was that the AL specialist should work with subject lecturers, and that

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

content-subject lecturers must be involved in the evaluation process; this would be particularly relevant in the case of discipline-specific AL interventions. These suggestions could be fruitful, as content-subject lecturers are likely to be able to give valuable input as to which instruments might measure the transfer of AL abilities the best.

Three final comments were made that are particularly valuable. Firstly, it was pointed out that *students' responses to questionnaires etc. would obviously be subjective and likely to be influenced by current academic performance so the interviews/questionnaires would need to be considered alongside their writing*. This highlights the need for triangulation in an evaluation design, and the necessity of including a variety of instruments. A second comment highlighted the need for the evaluation design to be flexible: *It can never be a one size fits all. My use of the design and approach of AL is that you have to be sensitive to context, to social practices and these are always dynamic, changing, and negotiated. In other words, evaluation has to also be flexible and depends on space, time, and purpose*. A final suggestion was that the evaluation process itself should ultimately be evaluated. Section 4 of this article attempted to do just that through a process of self-reflection, and it was indeed a valuable exercise, as I would approach future impact evaluations slightly differently, based on the challenges experienced in this evaluation process.

The following section proposes a revised evaluation design based on personal experience of implementing some of the instruments proposed in the initial design, as well as feedback from AL specialists across South Africa.

REVISED EVALUATION DESIGN

The revised evaluation design (Figure 3), the result of this process of negotiation, includes some added instruments, some revised instruments, as well as general guidelines for the implementation of the design.

Instruments that were added to the initially proposed evaluation design (Figure 1) are as follows:

- assessing students' academic speaking abilities (Appendix D),
- assessing students' listening abilities,
- using qualitative student data, and
- benchmarking with other institutions.

Instruments that were revised are:

- student need-press questionnaire on AL abilities (Appendix A),
- lecturer questionnaire on AL abilities (Appendix B),
- writing rubric (Appendix C), and
- individual extended written assignment (to also allow for group assignments or assignments from students' content subjects).

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

Figure 3: Revised evaluation design for academic literacy interventions.

The initially proposed evaluation design was further adapted in that the categories of 'recommended instruments' and 'additional optional instruments' were removed. As the quantitative responses in Section 5 indicate, those instruments that had originally been classified as 'recommended' are not necessarily seen as

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

more applicable to AL specialists responsible for specific categories of interventions than any of the other instruments. Note that each of the four categories of interventions contains the same instruments. However, instruments are ordered, from top to bottom for each respective category (indicated by different colours), by how relevant each instrument might be based on feedback from AL specialists (see Section 5). Note that this order is merely an estimation of usefulness, and the order of relevance of instruments will without doubt be different for each individual intervention within these four categories. It is also important to note that the colour-coded categories can overlap in some cases. For instance, a subject-specific AL test, or a writing assignment completed for another subject would indicate whether there was an improvement in students' AL levels (and would thus be categorized under the blue-coded research instruments), but could also provide information regarding the transfer of such abilities (thus, it might also be applicable under the green-coded categories) (see Figure 3).

The original recommendation that at least three instruments be used in cases where control groups are not available, and at least two instruments be used where control groups are available, remains unchanged (of course, the more instruments that are used, the more complete the picture that will emerge). So does the recommendation of using at least one instrument that indicates whether there was an improvement in students' AL levels (indicated in blue), and one instrument that indicates transfer of these abilities or the need for them in students' content subjects (indicated in green). Instruments that do not clearly fall into either of these categories, but that might still be useful, are indicated in orange. By using a number of instruments from various categories, the researcher strengthens triangulation by both source and method (cf. Lynch 1996). To further strengthen triangulation, it is suggested that a combination of qualitative and quantitative instruments be used wherever possible.

One of the key characteristics of this evaluation design for AL interventions is its flexibility. Evaluators must be able to choose instruments that are applicable to their respective contexts, and even interventions in the same category (e.g. generic AL courses) might need to use very different instruments than other interventions in the same category. The greater the variety of instruments used (e.g. qualitative and quantitative instruments, and instruments that measure an improvement in AL abilities as well as those that determine the necessity of abilities in students' content subjects and the possible transfer of these abilities), and the larger the number of instruments used, the more valid the deductions regarding impact and its causality that can be made based on the triangulated data obtained from these instruments, provided that the instruments used are sound, valid and of high quality. Using poorly designed instruments will drastically influence the validity of any evaluation study.

Even if a range of high quality and relevant instruments is used and data are triangulated, this design still has limitations. In the course that was evaluated in Fouché *et al.* (2017) and Fouché (2017), it seemed clear that students improved significantly in several AL abilities, that students believe that the areas they had shown improvement in were necessary for success in their other

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

subjects, and that these abilities were generally sufficiently addressed in their AL course. It would also seem as though most of the AL areas in which students had improved showed a significant improvement when content-subject assignments were assessed at the end of the year. Yet, there is still no definite empirical proof that the AL course itself was responsible for such improvement, despite strong indications from the literature that AL abilities are unlikely to improve without a specific intervention (see, e.g. Thompson 1990; Farnill and Hayes 1996; Rosenthal 1996; Holder *et al.* 1999; De Graaff and Housen 2009). However, based on the rich and varied evidence provided, it would seem likely that the AL course did have a meaningful impact.

Ultimately, it remains the researcher's duty to ensure that the most comprehensive combination of instruments that is feasible is used to strengthen inferences regarding the impact of the intervention as far as is possible, whilst acknowledging any limitations that might still remain in the research design. From a critical realist perspective, this means that the researcher works as responsibly as possible within the Empirical domain to attempt to describe relevant events in the Actual domain, whilst being cognizant of the fallibility of humans in their attempt to understand more of the Real domain. If data are responsibly triangulated by both source and method (cf. Lynch 1996), much value will be added to the field of AL by working towards effective interventions that have the highest impact possible for their specific contexts.

CONCLUSION

Kiely (2009: 99) calls for research that ensures that 'the research-type knowledge-building enterprise and the ongoing quality management processes are mutually informing, and that programme evaluation becomes a socially-situated cycle of enquiry, dialogue, and action'. Cole *et al.* (2005) echo this sentiment, and point out that this process is characterized by a generate-test cycle. The proposed evaluation design should therefore be seen as critical realist artefacts that can, and must be adapted over time in the complex interplay between manufacturer, product, and community of use, taking into consideration the intent of the former, the expectations of the latter (cf. Figure 2; Buchanan 2001), and the ultimate reality of the implemented artefact. The self-reflection and feedback from specialists discussed in this article formed part of this design research process, but it is a process that must be continued in future research for various contexts.

Cole *et al.* (2005) point out though that the generate-test cycle is constrained by available resources as well as technology—this might be why so few studies have attempted to comprehensively measure impact (Fouché 2015). Where such constraints exist, it might be wise to use existing instruments, albeit imperfect and fallible, from a critical realist perspective (cf. Lindén *et al.* 2017), and to adapt these where possible. It is certainly preferable to follow that route rather than to avoid measuring the impact of an intervention altogether since no perfect instruments exist with which to do this.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

The current article has indicated how research, in particular self-reflection as well as feedback from specialists in the field, can assist in solving either current or anticipated problems of practitioners (cf. Cole *et al.* 2005). By continuously integrating such reflection into the research process, the artefact's utility, quality, and efficacy (cf. Hevner *et al.* 2004; Cole *et al.* 2005) will continuously be improved upon, and the field as a whole will benefit.

REFERENCES

Alfaro-Tanco, J. A., et al. 2021. 'An evaluation framework for the dual contribution of action research: Opportunities and challenges in the field of operations management,' *International Journal of Qualitative Methods* 20: 16094069211017761–16.

Archer, A. 2008. 'Investigating the effect of Writing Centre interventions on student writing,' *South African Journal of Higher Education* 22/2: 248–64.

Ardington, C., et al. 2020. 'Technical report: Benchmarking early grade reading skills in Nguni languages'. ReSEP, Stellenbosch University.

Babbie, E. R. 2021. *The Practice of Social Research*. Cengage Learning.

Bachman, L. F. and **A. S. Palmer**. 2010. *Language Assessment in Practice*. Oxford University Press.

Bamberger, M., **J. Rugh**, and **L. Mabry**. 2012. *'Real World Evaluation: Working under Budget, Time, Data and Political Constraints,'* Sage.

Baum, F., **C. Macdougall**, and **D. Smith**. 2006. 'Participatory action research,' *Journal of Epidemiology and Community Health* 60/10: 854–7.

Bharuthram, S. and **S. Clarence**. 2015. 'Teaching academic reading as a disciplinary knowledge practice in higher education,' *South African Journal of Higher Education* 29/2: 42–55.

Bhaskar, R. 2008. *Dialectic: The pulse of freedom*. Routledge.

Boakye, N. and **M. Mai**. 2016. 'A needs analysis for a discipline-specific reading intervention,' *English Language Teaching* 9/3: 235–47.

Buchanan, R. 2001. 'Design research and the new learning,' *Design Issues* 17/4: 3–23.

Butler, G. 2009. 'The design of a postgraduate test of academic literacy: Accommodating student and supervisor perceptions,' *Southern African Linguistics and Applied Language Studies* 27/3: 291–300.

Carstens, A. 2009. *The effectiveness of genre-based approaches in teaching academic writing: Subject-specific versus cross-disciplinary emphases*. Unpublished PhD dissertation, University of Pretoria.

Carstens, A. and **A. Rambiritich**. 2021. 'Directiveness in tutor talk,' *Perspectives in Education* 39/3: 151–68.

Cattani, G., **S. Ferriani**, and **P. D. Allison**. 2014. 'Insiders, outsiders, and the struggle for consecration in cultural fields: A core-periphery perspective,' *American Sociological Review* 79/2: 258–81.

Chawla-Duggan, R. 2007. 'Breaking out, breaking through: Accessing knowledge in a non-western overseas educational setting—methodological issues for an outsider,' *Compare* 37/2: 185–200.

Clarence, S. and **S. McKenna**. 2017. 'Developing academic literacies through understanding the nature of disciplinary knowledge,' *London Review of Education* 15/2: 38–49.

Cole, R., et al. 2005. 'Being proactive: Where action research meets design research' in ICIS International Conference on Information Systems, Las Vegas, USA, December 11–14, Association for Information Systems (AIS), pp. 325–335. https://aisel.aisnet.org/icis2005/27/.

Connor, M. J. 2004. 'The practical discourse in philosophy and nursing: an exploration of linkages and shifts in the evolution of praxis,' *Nursing Philosophy* 5/1: 54–66.

Corson, D. 1997. 'Critical realism: An emancipatory philosophy for Applied Linguistics?,' *Applied Linguistics* 18/2: 166–88.

De Graaff, R. and **A. Housen**. 2009. 'Investigating the effects and effectiveness of L2 instruction' in **Long, M.H.** and **Doughty, C.J.** (eds.): *The Handbook of Language Teaching*. Wiley-Blackwell, pp. 736–755.

De Vos, A. S., et al. 2011. *Research at Grassroots for the Social Sciences and Human Service Professions*. Van Schaik.

Dewey, J. 1933. 'Analysis of reflective thinking' in **Hickman, L.A.** and **Alexander, T.M.** (eds.): *Reprinted in 1998 in The Essential Dewey*. Indiana Unviersity Press, pp. 137–150.

Dillman, D.A. 2007. *Mail and Internet Surveys: The Tailored Design Method*. John Wiley & Sons.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

Dison, L. and **B. Mendelowitz**. 2017. 'Reflecting centre students' of writing a contextualised practice centre experiences' in **Clarence, S.** and **Dison, L.** (eds.): *Writing Centres in Higher Education*. SUN Press, pp. 93–208.

Dison, L., et al. 2022. 'Reframing purpose and conceptions of success for a post-Covid-19 South African higher education,' *Scholarship of Teaching and Learning in the South* 6/1: 33–54.

Farnill, D. and **S. Hayes**. 1996. 'Do NESB university students with poor English skills make rapid linguistic gains in mainstream studies?,' *Higher Education Research and Development* 15/2: 261–8.

Fouché, I. 2015. 'Towards impact measurement: An overview of approaches for assessing the impact of academic literacy abilities,' *Stellenbosch Papers in Linguistics* 44/1: 19–35.

Fouché, I. and **S. Immelman**. 2015. 'Rubric for assessing oral presentations,' *LST 143 Workbook*. University of Pretoria.

Fouché, I., **T. Van Dyk**, and **G. Butler**. 2016. 'Impact measurement: towards creating a flexible evaluation design for academic literacy interventions,' *Stellenbosch Papers in Linguistics* 45/1: 109–45.

Fouché, I. 2017. 'Impact measurement: quantitatively determining the improvement in students' academic literacy levels at a South African university,' *Journal for Language Teaching* 51/1: 163–99.

Fouché, I., **T. Van Dyk**, and **G. Butler**. 2017. 'An "enlightening course that empowers first years"?: A holistic assessment of the impact of a first-year academic literacy course,' *Journal of English for Academic Purposes* 27: 14–30.

Gasman, M. and **L. Payton-Stewart**. 2006. 'Twice removed: A white scholar studies the history of black sororities and a black scholar responds,' *International Journal of Research & Method in Education* 29/2: 129–49.

Gee, J. 2008. *Social Linguistics and Literacies*. Routledge.

Granott, N. 1998. 'We Learn, Therefore We Develop: Learning Versus Development—or Developing Learning?' in **Smith, M.C.** and **Pourchot, T.** (eds.): *Adult Learning and Development: Perspectives from Educational sychology*. Cambridge University Press, pp. 213–242.

Hawkey, R. 2006. *Studies in language testing: Impact Theory and Practice*. Cambridge University Press.

Hevner, A. and **S. Chatterjee**. 2010. 'Design science research in information systems' *Design Research in Information Systems*. Springer, pp. 9–22.

Hevner, A. R., et al. 2004. 'Design science in information systems research,' *MIS quarterly* 28/1: 75–105.

Holder, G. M., et al. 1999. 'Academic literacy skills and progression rates amongst pharmacy students,' *Higher Education Research & Development* 18/1: 19–30.

Jacobs, C. 2005. 'On being an insider on the outside: New spaces for integrating academic literacies,' *Teaching in Higher Education* 10/4: 475–87.

Jick, T. D. 1979. 'Mixing qualitative and quantitative methods: Triangulation in action,' *Administrative Science Quarterly* 24/4: 602–11.

Judd, T. and **B. Keith**. 2017. 'Implementing undergraduate student learning outcomes assessment at the program and institutional levels' *Handbook on Measurement, Assessment, and Evaluation in Higher Education*. Routledge, pp. 69–86.

Kerstetter, K. 2012. 'Insider, outsider, or somewhere between: The impact of researchers' identities on the community-based research process,' *Journal of Rural Social Sciences* 27/2: 7.

Kiely, R. 2009. 'Small answers to the big question: Learning from language programme evaluation,' *Language Teaching Research* 13/1: 99–116.

Lea, M. R. 2004. 'Academic literacies: a pedagogy for course design,' *Studies in Higher Education* 29/6: 739–756.

Lea, M. R. and **B. V Street**. 1998. 'Student writing in higher education: An academic literacies approach,' *Studies in Higher Education* 23/2: 157–72.

Lehohla, P. 2012. Census in brief 2011. Statistics South Africa. https://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf.

Liem, A. 2011. 'Using design education to survive in the'corporate world'of higher learning and research,' *Journal of Design Research* 9/2: 104–18.

Lillis, T. 2003. 'Student writing as' Academic Literacies': Drawing on Bakhtin to move from critique to design,' *Language and Education* 17/3: 192–207.

Lindén, J., **J. Annala**, and **K. Coate**. 2017. 'The role of curriculum theory in contemporary higher education research and practice,' *Theory and Method in Higher Education Research* 3: 137–54.

Lynch, B. K. 1996. *Language Program Evaluation: Theory and Practice*. Cambridge University Press.

Lynch, B. K. 2003. *Language Assessment and Programme Evaluation*. Edinburgh University Press.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023

Mandviwalla, M. 2015. 'Generating and justifying design theory,' *Journal of the Association for Information Systems* 16/5: 3143–344.

Marais, F. and **T. Van Dyk**. 2010. 'Put listening to the test: An aid to decision making in language placement,' *Per Linguam* 26/2: 1–19.

Meyers, G. L., **M. Jacobsen**, and **E. Henderson**. 2018. 'Design-Based Research: Introducing an innovative research methodology to infection prevention and control',' *Canadian Journal of Infection Control* 33/3: 158–164.

Mhlongo, G. J. 2014. *The impact of an academic literacy intervention on the academic literacy levels of first year students: The NWU (Vaal Triangle Campus) experience*. Unpublished MA dissertation, North West University.

Michaelsen, L. K. and **M. Sweet**. 2011. 'Team-based learning,' *New Directions for Teaching and Learning* 2011/128: 41–51.

Muhammad, M., et al. 2015. 'Reflections on researcher identity and power: The impact of positionality on community based participatory research (CBPR) processes and outcomes,' *Critical Sociology* 41/7-8: 1045–63.

Ngwenya, T. 2010. 'Correlating first-year law students' profile with the language demands of their content subjects,' *Per Linguam* 26/1: 74–99.

Pienaar, C. 2005. 'Shared assessment: Empowering student writers,' *Language Matters* 36/2: 193–204.

Rambiritch, A. 2013. 'Validating the Test of Academic Literacy for Postgraduate Students (TALPS),' *Journal for Language Teaching* 47/1: 175–93.

Rosenthal, J. W. 1996. *Teaching Science to Language Minority Students: Theory and practice*. Multilingual Matters.

Saleh, A. and **K. Bista**. 2017. 'Examining factors impacting online survey response rates in educational research: Perceptions of graduate students,' *Journal of MultiDisciplinary Evaluation* 13/29: 63–74.

Scott, D. 2005. 'Critical realism and empirical research methods in education,' *Journal of Philosophy of Education* 39/4: 633–46.

Simon, H. A. 1996. *The Sciences of the Artificial*, 3rd ed. MIT Press.

Thompson, R. M. 1990. 'Writing-proficiency tests and remediation: Some cultural differences,' *TESOL Quarterly* 24/1: 99–102.

Van De Poel, K. and **T. Van Dyk**. 2015. 'Discipline-specific academic literacy and academic integration' in **Wilkinson, R.** and **Walsh, M.L.** (eds.): *Integrating Content and Language in Higher Education*. Peter Lang, pp. 161–180.

Van Dyk, T. and **K. Van De Poel**. 2013. 'Towards a responsible agenda for academic literacy development: Considerations that will benefit students and society,' *Journal for Language Teaching* 47/2: 43–69.

Van Dyk, T., et al. 2009. 'On being reflective practitioners: The evaluation of a writing module for first-year students in the Health Sciences,' *Southern African Linguistics and Applied Language Studies* 27/3: 333–44.

Van Dyk, T., et al. 2011. 'Ondersoek na die impak van 'n akademiese geletterdheidsintervensie op eerstejaarstudente se akademiese taalvermoë,' *LitNet Akademies* 8/3: 487–506.

Van Rooy, B. and **S. Coetzee-Van Rooy**. 2015. 'The language issue and academic performance at a South African university,' *Southern African Linguistics and Applied Language Studies* 33/1: 31–46.

Van Wyk, A. 2014. 'English-medium education in a multilingual setting: A case in South Africa,' *International Review of Applied Linguistics in Language Teaching* 52/2: 205–20.

Weigle, S. 2002. *Assessing Writing*. Cambridge University Press.

Winberg, C., et al. 2013. 'Conceptualising linguistic access to knowledge as interdisciplinary collaboration,' *Journal for Language Teaching* 47/2: 89–107.

Wingate, U. 2015. *Academic Literacy and Student Diversity*. Multilingual Matters.

Yamashita, J. 2002. 'Mutual compensation between L1 reading ability and L2 language proficiency in L2 reading comprehension,' *Journal of Research in Reading* 25/1: 81–95.

Yost, D. S., **S. M. Sentner**, and **A. Forlenza-Bailey**. 2000. 'An examination of the construct of critical reflection: Implications for teacher education programming in the 21st century,' *Journal of teacher education* 51/1: 39–49.

Downloaded from https://academic.oup.com/applij/article/44/4/718/7169413 by University of Witwatersrand user on 11 September 2023