Introduction

Measuring vocabulary learning gains is a key aspect of language teaching as well as of multiple research studies on language learning. In pedagogy, vocabulary tests are often used to check lexical learning progress summatively at the end of one or several instructional units. The purpose of such tests is to check how much of the taught vocabulary has been acquired by language learners, and thus indirectly to check the quality of the instruction, or inform future learning activities. In research, vocabulary gains are often measured to investigate the effectiveness of a particular teaching or learning method. Numerous studies have thereby looked at various aspects of vocabulary acquisition. Studies have been conducted to examine incidental vocabulary learning, from written input through extensive or narrow reading (e.g., Barcroft, 2009; Fraser, 1999; Horst, 2005; Horst, Cobb, & Meara, 1998; Hwang & Nation, 1989; Paribakht & Wesche, 1993; Pellicer-Sánchez, 2016; Pellicer-Sánchez & Schmitt, 2010; Pigada & Schmitt, 2006; Rott, 1999; Saragi, Nation, & Meister, 1978; Waring & Takaki, 2003; Webb & Chang, 2015), as well as from oral or audio-visual input through listening to conversations, lectures, radio, or stories (e.g., Elley, 1989; Vidal, 2003, 2011; van Zeeland & Schmitt, 2013), or viewing films and television (Montero Perez, Peters, Clarebout, & Desmet, 2014; Webb, 2015), or even a mixture of modalities (e.g., Brown, Waring, & Donkaewbua, 2008; Pellicer-Sánchez et al., 2018). Other studies have probed differences between incidental and intentional vocabulary learning, and have compared which approach leads to greater vocabulary learning progress (e.g., Min, 2008). A strand of research similar to this has investigated the effectiveness of different vocabulary teaching and learning activities or enhancement techniques in vocabulary learning (Bauer & Nation, 1993; Beaton, Gruneberg, & Ellis, 1995; Ellis & He, 1999; Hulstijn, 1992; Hulstijn & Trompetter, 1998; Hulstijn, Hollander, & Greidanus, 1996; Joe, 1998; Knight, 1994; Laufer, 2000; Newton, 1995; Paribakht & Wesche, 1997; Watanabe, 1997). Comparing measured vocabulary learning progress has also been employed to research the beneficence of various learning conditions, such as elaboration effects (e.g., Ellis & He, 1999), repetition effects (e.g., Horst et al., 1998; Webb, 2007a; Webb & Chang, 2012; Webb, Newton, & Chang, 2013), the effects of spaced rehearsal (Nakata, 2015), or study conditions (Nakata & Webb, 2016), among many more. While most of these studies
have focused on vocabulary learning gains in the form-meaning link knowledge dimension, i.e., how effective various approaches were for acquiring the meaning of an L2 form, a number of researchers have also investigated the learning of other word knowledge aspects, such as collocations (e.g., Boers, Demecheleer, Coxhead, & Webb, 2014; Boers, Demecheleer, & Eyckmans, 2004; Laufer & Girsai, 2008; Lindstromberg & Boers, 2008; Webb & Kagimoto, 2009, 2011; Webb et al., 2013).

The measurement of vocabulary gains is therefore probably the most common type of measurement, both in language classrooms as well as in vocabulary research studies. Any intervention study will have to provide robust measurement to demonstrate whether a particular treatment was or was not effective. However, this also seems one of the most challenging types of measurement as there are a number of critical issues that require careful consideration. This chapter will discuss three of these in the following section.

Critical Issues and Topics

Study Design

In research, examining vocabulary gains often takes the form of intervention studies. Such studies involve one or more experimental treatments of one or more groups of learners and measurements of vocabulary knowledge before and after this intervention. The measurement of vocabulary gains implicitly presupposes some kind of baseline that indicates the vocabulary knowledge before any type of learning would occur. This baseline is often referred to as the pretest. The pretest determines the level of knowledge before the treatment and is crucial to know how much room for learning there is for the learners and which vocabulary items (and which word knowledge aspects of these) are already known by the learners.

The intervention or treatment may then take various forms. Depending on whether incidental or intentional vocabulary learning is to be studied, the treatment can range from giving students reading, listening or viewing materials, to having them work through a number of tasks or exercises, or providing them with a particular method of instruction. Often, this also involves some kind of comparison, so that two or more groups of learners receive different treatments, or that one or more groups receive a treatment (i.e., experimental groups) and one group does not (i.e., control group). Ideally, we should make sure that these different groups are as similar as possible in terms of their background variables, proficiency, and, most importantly, their vocabulary knowledge in terms of the target knowledge to be investigated, which is again indicated by the scores on the pretest. One way to ensure this is to assign learners at random to one of the examined groups. This, however, is often challenging when working with intact school classes. If the learner groups differ considerably in their pretest scores, then this should be accounted for when analyzing and interpreting the posttest scores.

After the intervention, there needs to be some type of posttest to identify the vocabulary gains. This test needs to be identical to the one administered prior to the treatment. Only through this comparison of knowledge in the pretest vs. posttest can we know how much or what about the vocabulary items has or has not been learned. If a different test is used as a posttest than was used as a pretest, we cannot be certain that any difference in scores from the two administrations are, in fact, due to vocabulary learning from the intervention or input, and not due to the nature of the different tests.

There are two different types of posttests. An immediate posttest is administered at the end of an intervention and measures the immediate gains of the treatment. A delayed posttest
may be administered several days or weeks after the immediate posttest or treatment. A delayed posttest is generally employed when we are interested in longer-term retention of vocabulary gains. Researchers have to decide whether they want to administer either an immediate or a delayed posttest, or both. Only using an immediate posttest in a research design, however, often results in an overestimation of the vocabulary learning that has actually taken place (Waring & Takaki, 2003). Rott (1999) has rightly criticized the lack of a delayed posttest as a measurement of long-term learning effects as one of the key shortcomings of much research on incidental word learning from reading, among other factors, such as failure to employ control groups and using measuring instruments that lack the necessary sensitivity to pick up on incremental changes in word knowledge.

Posttests will usually show an effect immediately after the treatment, but robust learning may not be found after some time has passed. Schmitt (2010) therefore argues for the need for a delayed posttest in vocabulary learning studies. He states that the information gleaned from immediate posttests is limited because it disregards the fact that learning is “always liable to attrition” (Schmitt, 2010, p. 155) and only gives a snapshot view of vocabulary knowledge when we know the learning process is incremental, dynamic, and nonlinear. This leads to limited generalizability of immediate posttest scores. In addition, Schmitt (2010) states that initial practice will almost always be more effective than later or additional practice, where returns will be diminishing. Using only an immediate posttest would therefore again be overestimating practice effects and thus durable learning. In other words, while an immediate posttest gives an indication of the effectiveness of the treatment, only a delayed posttest should be interpreted as an indication of learning (Schmitt, 2010).

Nation and Webb (2011) outline some further crucial points to consider when conducting intervention studies with pretests and posttests. First, all tests need not only be identical but also need to be administered identically. Moreover, researchers need to be aware that a pretest may already prime test takers to certain learning, which may confound results. For instance, in a study on incidental vocabulary learning, a pretest may alert learners to pay more attention to the lexical items from the pretest and thereby bias findings. Nation and Webb (2011) suggest that this could be avoided by using filler or distracter items both in the test as well as the intervention, i.e., including items in the test that will not appear in the experiment and vice versa, so that learners will not be unduly influenced by the pretest. Researchers also need to be aware of the fact that the testing event itself constitutes exposure and a learning event. Pretests and posttests “involve motivated deliberate attention to the tested words” (Nation & Webb, 2011, p. 277), which may result in learning. Control groups who only take the pretest and posttest, or splitting the experimental groups to have them take either the immediate or the delayed or both posttests, could again be a solution to account for testing effects (Avila & Sadoski, 1996). This, however, is sometimes unfeasible given limited sample sizes. Further, researchers are often unable to control whether any deliberate (or even incidental) learning of participants takes place between an immediate and a delayed posttest. Particularly with small groups of test takers, only a few learners engaging in deliberate learning may affect the outcomes considerably (Nation & Webb, 2011). On the other hand, a delayed posttest may enable an investigation of exactly such intervening variables that may influence retention (Nation & Webb, 2011).

The ideal research design for measuring vocabulary gains after one particular treatment may therefore look as shown in Table 26.1.

A control group takes all tests but does not receive any treatment. Two experimental groups receive the same treatment, but one of them only sits the pretest and the delayed
posttest, while the other group completes all tests. This then allows for singling out the effect of the treatment as well as the effect of the immediate posttest.

With a comparison of different treatments, this design would technically inflate to involving five different groups (see Table 26.2). This is often impractical, so researchers may revert to only using two experimental groups (e.g., Experimental I and III).

When measuring vocabulary gains in classroom situations, teachers often only have one experimental group and no control group. This usually suffices for small-scale interest of assessing the achievement of learners but does not allow for comparisons to other (or no) interventions. Any inferences about students’ learning still requires the basic pretest and posttest design.

A further key issue in implementing this design is the timing of the delayed posttest. There is no standardized recommendation for how delayed a delayed posttest should be. The interval between the end of the treatment and the delayed posttest usually spans between one to two weeks, sometimes even extending to four weeks or several months. Schmitt (2010), for example, recommends a span of three weeks and no less than one week. However, the span is often based on a theoretical convention rather than an empirically motivated rationale. Nation and Webb (2011) thus rightly caution that the strength of the treatment effect and the sensitivity of the measurement instrument need to be considered when deciding on the delayed posttest interval. They recommend that the interval should be determined based on the context and purpose, depending on how long one would “want and expect learners to retain their knowledge of the word gained from the treatment” (Nation & Webb, 2011, p. 280). For instance, in a classroom-based study, this interval might be determined by looking at when the learners would be likely to encounter the learned word again, which would often mean a minimum interval of at least two to three days between the immediate and the delayed posttest as this would be a typical interval between two lessons. Other researchers have stated that a two week interval would be more appropriate as teaching materials or graded readers typically do not repeat new words sooner than that (Nation & Wang, 1999).

Research has shown that sleep consolidation may also have some effect in this and needs to

### Table 26.1 Model research design for investigating vocabulary learning progress after a treatment

<table>
<thead>
<tr>
<th>Group</th>
<th>Pretest</th>
<th>Treatment</th>
<th>Immediate posttest</th>
<th>Delayed posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental I</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental II</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Table 26.2 Model research design for investigating vocabulary learning progress through different treatments

<table>
<thead>
<tr>
<th>Group</th>
<th>Pretest</th>
<th>Treatment A</th>
<th>Treatment B</th>
<th>Immediate posttest</th>
<th>Delayed posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental I</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental II</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental III</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental IV</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
be taken into account when deciding on the delay. It has been found that sleep sometimes even increases the scores on a posttest that is only delayed by one day (Davis, Di Betta, Macdonald, & Gaskell, 2008; Dumay & Gaskell, 2007; Gaskell & Dumay, 2003). The appropriate delay for any particular measurement will therefore have to be decided in consideration of such context-dependent variables.

There are a few exceptions to the design outlined earlier that would still allow measuring vocabulary learning progress. One such exception is when the immediate posttest is treated as another form of pretest, as is sometimes done in vocabulary attrition research. In attrition studies, researchers will often want to take as a starting point the time when all words were demonstrated as known on some test. This may involve repeated measurements with immediate posttests after learning activities to determine at what point all lexical items were known and thus the attrition process may set in. Repeated immediate posttests thereby also allow investigating how much learning effort might have gone into mastering a particular item; the more immediate posttests required to show that an item is learned, the more difficult that item is to learn. The ultimate immediate posttest trial could be treated as a starting point and thus a pretest before other delayed posttests – with no treatments in between – will be administered.

Another such exception is when non-words are used in a study instead of real words. Since it is fair to assume that participants in the intervention study will not be familiar with words that the researcher made up, no pretest will be necessary to establish the nonexistent previous knowledge of the target words. While such a design may have the disadvantage of not having much real-life practical use for participants and may potentially be lacking in ecological validity, it is a very effective way to eliminate any priming effects that may occur through the use of a pretest. However, since the non-words are often mapped onto meanings of real words, and researchers may not control for knowledge of those in the participants’ L1 or L2, there may also be some bias in findings of such studies. However, Webb (2007b) states that the use of non-words, or disguised forms of low-frequency words, has advantages because it is often difficult to find target words that no participant has any prior knowledge of. It seems unclear, though, whether this problem does not also hold for the low-frequency words or concepts that form the basis of the disguised forms. He further states that non-words or disguised form’s lack of requiring a pretest is beneficial because pretests are not typically sensitive enough to pick up prior partial knowledge of items. In any case, this non-word approach has been successfully employed in a number of studies (Pellicer-Sánchez, 2016; Webb, 2007b).

Another approach that circumvents the use of formal pretests and posttests is sometimes found in psycholinguistic studies of incidental learning. In studies such as that by Pellicer-Sánchez (2016), initial reading times of new words may be compared to the reading times of those words after repeated exposure. In such designs, the initial reading time might be considered the pretest, and the final reading time the immediate posttest, while the repeated exposures would represent a treatment of sorts leading to an improvement in processing time and thus an indication of vocabulary learning gains. In most designs, however, vocabulary learning gains are measured more explicitly with the use of formal tests. The next section will problematize some issues with such tests.

**Instruments**

Much like there is no standard period of delay for posttests, there is also no standard instrument to measure vocabulary gains. The choice of test, or rather test format, usually depends on what aspect of word knowledge learning is intended to be captured. While in most cases,
Measuring Vocabulary Learning Progress

measuring vocabulary learning will be taken to mean measuring the number of words for which a previously unknown meaning has been acquired, it is possible to conceive of interventions that focus on other aspects of word knowledge, such as pronunciation or spelling, in which case a specific test or test format would need to be used to assess these knowledge aspects before and after the learning period. Nation and Webb (2011) recommend using several measures of vocabulary knowledge in any case, as there may be incidental learning of various (even unintended) word knowledge aspects taking place during a treatment. Even for the same word knowledge aspect, it might be advisable to use several different tests or test formats. For instance, a battery of recall and recognition tests may be used for just the form-meaning link knowledge aspect to determine the strength of the learning. This is particularly important as one instrument alone is often not sensitive enough to detect minimal changes in the incremental learning of vocabulary knowledge aspects. Particularly standardized vocabulary tests, such as the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001), the Vocabulary Size Test (Nation & Beglar, 2007), or the updated Vocabulary Levels Test (Webb, Sasao, & Ballance, 2017), are not designed or intended to be used for pretest and posttest designs with short-term interventions. These measures of vocabulary breadth are not fine-grained enough to pick up small changes in vocabulary size, even when supposedly parallel versions of the tests are used. While general vocabulary size is unlikely to increase noticeably over short periods of instruction, the test formats they employ, i.e., Multiple Choice and Multiple Matching for form or meaning recognition, may still be feasible when targeting specific words that are focused on in the treatment. The choice of test format will mainly depend on the degree of sensitivity useful for investigating the gains. One format that is often credited for being more sensitive to vocabulary learning gains than other formats is the Vocabulary Knowledge Scale (VKS) (Wesche & Paribakht, 1996).

Vocabulary Knowledge Scale (VKS)

The VKS was designed to track small incremental changes in vocabulary knowledge as it follows a developmental approach to vocabulary measurement (Paribakht & Wesche, 1997). In its original version, it asks learners to indicate their degree of word knowledge on a five-point elicitation scale ranging from “I don’t remember having seen this word before” to “I can use this word in a sentence” (Paribakht & Wesche, 1993).

I I don’t remember having seen this word before.
II I have seen this word before, but I don’t know what it means.
III I have seen this word before, and I think it means _______ (synonym or translation).
IV I know this word it means ________ (synonym or translation).
V I can use this word in a sentence: _________ (write a sentence).

The scale can thereby be employed for every L2 target word, and learners, after seeing the decontextualized target, have to report the stage of mastery for each target. Stages III, IV, and V are complemented by a request to provide a synonym, L1 translation, or sentence. The original authors thereby claim that the VKS combines “self-report and performance items” (Paribakht & Wesche, 1997). The scale is popular for intervention study designs because it takes into account the partial and incremental process of word knowledge and combines assessing the form-meaning link as well as some aspects of depth of word knowledge. While Laufer and Goldstein (2004) argue that the VKS is an indirect test of word meaning rather than a test of vocabulary depth, it is often classified as the latter.
The VKS surpasses ordinary self-report scales in that the higher stages of the scale require not only a self-assessment but also demonstration of that self-asserted knowledge. The major problem, however, with the VKS, as with all other developmental approaches to vocabulary measurement, is that we do not know enough about the incremental acquisition of vocabulary knowledge in order to decide on the best scale to measure it (Schmitt, 2010). Even Wesche and Paribakht (1996) acknowledge the “lack of theoretical consensus about the nature and course of development of L2 vocabulary knowledge” (p. 32) and therefore the arbitrariness of the VKS. Neither the number of levels or stages nor the actual stages are grounded in a sufficient amount of empirical research to design a measurement scale which would allow for highly valid claims. Also, Schmitt (2010) argues that the scale might not be unidimensional in that it involves a “constellation of lexical knowledge” (p. 220) at the different stages, mixing receptive and productive elements in an unprincipled way and offering various degrees of contextualization. Schmitt (2010) further echoes Read’s (2000) critique that the intervals between the five stages might not be equidistant. Stewart, Batty, and Bovee (2012) explored the psychometric dimensionality of the VKS empirically and found a weak multidimensionality and unclear construct distinctions. The close difficulty proximity of some knowledge levels, they argue, impedes the results’ interpretability and the VKS’ usefulness as a diagnostic measure for educators.

In terms of demonstration of knowledge, produced sentences at the highest level, presumably showing the highest degree of mastery, bear a number of scoring issues as acceptable and even sophisticated sentences could be produced by candidates that do not sufficiently demonstrate knowledge of the target word (McNeill, 1996; Schmitt, 2010). The VKS does not come with adequate scoring rubrics and guidelines that would minimize marker subjectivity at this stage (Bruton, 2009). Bruton (2009) also argues that the VKS precludes L2 form recall and that in cases of homographs it is not clear for the candidate which core meaning is actually targeted. Schmitt and Zimmerman’s (2002) simplified variation of the VKS, which comprises four rather than five stages, unfortunately suffers from the same limitations in principle.

Despite this, the VKS has frequently been used as a research tool (e.g., Bruton, 2009; de la Fuente, 2002; Horst, Cobb, & Nicolae, 2005; Paribakht & Wesche, 1997; Paribakht, 2005; Pulido, 2004; Rott, Williams, & Cameron, 2002; Wesche & Paribakht, 2009). Golonka et al. (2015) even claim that the VKS is “the most widely used scale for measuring vocabulary depth” (p. 25). Despite its operationalization of some valid assumptions, however, its merit for measuring vocabulary learning gains is still contested. While it may be useful as a supplementary instrument for classroom teachers, particularly for capturing initial stages in word learning (Schmitt, 2010), Wesche and Paribakht (1996) themselves admit that the VKS “is not suitable for testing large numbers of students in its present form” (p. 33).

Given that available, standardized tests will rarely fit the purpose of any individual scenario in which we are interested in the vocabulary gains of learners, the majority of studies investigating vocabulary growth have designed bespoke measures for their respective purposes. While this in itself is not problematic, few of these research endeavors have concerned themselves with the need for validating such new, custom-made measurement tools. This seems potentially troublesome as any new test or any change to a vocabulary test would ideally require robust information on the functioning of its items before it is being used in research, so that findings can be interpreted with confidence. However, it is often not feasible to subject new instruments to the extensive piloting and test validation procedures that would be necessary. In that case, the reporting of descriptive statistics from the actual
administrations is arguably the bare minimum to demonstrate trustworthiness of the measurement instruments and thus the yielded results.

**Target Word Selection**

In light of most measurement tools being custom-made, another key consideration when measuring vocabulary learning progress is that of target word selection. When assessing learning gains in classroom settings, this selection will be driven by the syllabus or course book, i.e., the words that are supposed to be useful for students to learn and have their knowledge assessed on. In research scenarios, careful attention needs to be paid to choosing words that participants will not know at the time of the pretest. Studies measuring vocabulary learning gains need to allow for maximum learning possibilities at the outset. As mentioned before, non-words have been used to preempt potential existing knowledge of words before learning or teaching interventions in research settings. This ensures that no items are known before the intervention. However, aside from the issues raised already in the preceding section, there are also ethical issues to having candidates learn or study non-words that will not be of use to them outside the research setting.

Related to this is also the number of words that are selected for the pretest and post-test measurement. Fewer target words in both the intervention and the tests allow for more focused instruction or learning and potentially larger gains to be demonstrated. However, more target words could hypothetically give more learning opportunities, and also allow for more progress. Moreover, in incidental learning studies, filler items in the tests that are not the focus of the intervention may need to be considered, so as to mask which items will be targeted in the intervention. Practical considerations such as testing duration and participant fatigue also have to be weighed up when it comes to the item selection. For example, it may be possible to measure one word knowledge aspect of many words, or several word knowledge aspects for fewer vocabulary items. However, it may be challenging to measure several aspects of word knowledge for many words. Therefore, the limited number of words that most vocabulary learning intervention studies employ as their sample size, does indeed often make it challenging to generalize from findings to greater theoretical or pedagogical implications.

**Future Directions**

Three key issues seem to warrant particular attention by researchers in this area in the future. The first is the need for validating instruments thoroughly for making any substantial claims. Measurement instrument design or selection require deliberation and demonstration that any tool is assessing what is intended to be assessed validly and reliably to generate meaningful claims. To address this, one suggestion would be to encourage more collaboration between SLA vocabulary researchers and language testers, or for vocabulary researchers to engage more with measurement issues and the literature of the field of language testing (Kremmel & Pellicer-Sánchez, forthcoming). Ideally the field of vocabulary learning research would follow the model of Révész’ study (2012), which – albeit not a study on vocabulary learning – has been pointed out as exemplary in demonstrating validity evidence for research instruments within an SLA paper. Within her paper on the role of working memory and recasts on learning as measured by different instruments such as a grammaticality judgment task (GJT), a written picture description task, and two oral production tasks, she included “several analyses that could be considered validation evidence” (Purpura, Brown, & Schoonen, 2015, p.
for her outcome measurement instruments. Even though it was not systematically framed in an established validation approach, Révész provided different kinds of evidence that could be linked to different aspects or claims of current validation frameworks, demonstrating the soundness and validity of her claims based on results. For instance, she administered her tasks to different L1 groups to provide domain description evidence and ensure that the “tasks generated approximately the same number of obligatory contexts for the target form [past progressive], and that the use of the target construction was natural in each task” (2012, p. 106). She also examined the functionality of the developmental rating scale in detail using Multi-Facet Rasch Measurement (MFRM), which – in validation terms – contributed to her evaluation claims based on her instruments. Generalizability claims were supported in her study by, among other things, having a substantial portion of the data rated by two raters and providing a high inter-rater agreement value. Révész also investigated the model fit of her data in MFRM and found “expectedly high correlations between the productive skills and moderate correlations with the GJT, suggesting a meaningful relationship between knowledge of the past progressive (grammatical knowledge) and the ability to use this form in productive language use contexts (grammatical ability)” (Purpura et al., 2015, p. 53). This added backing that would support her explanation claims in an argument-based validation model.

She also examined whether the type of outcome measure influenced the learning of the past progressive (the feature in focus in her study) when recasts were employed, so that the measurement of learning would not have to rely on one source of measurement or information only. Emulating Révész’ study, therefore, there are numerous pieces of validation evidence about the tests and measurement instruments employed in a particular study that could and should be provided within papers on language (including vocabulary) learning. Adopting this approach is also what Purpura et al. (2015) suggest, alongside several other guidelines to improve the quality of quantitative SLA (vocabulary) research. Their recommendation to look to language test validation research and models as well as their checklist on measurement issues to consider (e.g., using a sufficient number of good quality test items, reporting descriptive statistics, appropriate reliability estimates, as well as details of how codings and raw scores were generated, checking assumptions and analyzing items thoroughly, describing the construct(s) in question, and selecting/designing appropriate measures to assess the construct(s) in focus) are important and overdue pointers for the future of the field.

The second key area for the future of research on vocabulary learning progress and its measurement is the issue of replication. The replicability of results is closely connected to the description of methodologies and measurement tools. If the latter is lacking, then it will be very difficult to replicate studies, which is direly needed in vocabulary learning research. This is particularly relevant, as recent research syntheses and meta-analyses have yielded that the reporting practices in the field are in need of improvement, even for the most basic information on the functioning of instruments or other key methodological aspects (e.g., Elgort, 2018; Huang, Eslami, & Willson, 2012). Most studies that have measured vocabulary learning progress have been one-offs, and have yet to be replicated robustly with controlled modifications of selected key variables, such as the L1 of participants, or participants’ proficiency levels. Sharing study data and materials openly on platforms such as the IRIS depository (Marsden, Mackey, & Plonsky, 2016) in the spirit of open scientific practices could be key to this endeavor.

A third area of particular interest in the future has already been opened up through technological advances. Online measures, such as eye tracking, are being increasingly used in SLA vocabulary research, and will allow us to investigate or measure not just the product of vocabulary learning progress, or regress for that matter, but shed light on the process and
cognitive factors influencing the outcome of vocabulary learning in increasing detail. They will also allow to exert much more control in learning experiments and will likely create exciting new avenues for the measurement of vocabulary learning progress.

Further Reading


Rott’s study offers a model approach of a research design that aims at measuring vocabulary learning progress, and offers a useful critique of previous studies on incidental vocabulary learning from reading.


Nation and Webb’s research handbook is a helpful practical guide for setting up measurements of vocabulary learning gains and discusses several considerations that need to be taken into account when designing an intervention study with pretests and posttests.


This paper by three key figures in SLA research and language assessment highlights the need for SLA researchers to look to language testing literature for support in validating their instruments. Since most SLA researchers interested in vocabulary learning engage in some form of measurement of language ability or knowledge, this is a useful overview of why it is essential that claims are based on validly and reliably operationalized variables in tests, instruments, and scores. It includes an exemplification of such an approach applied and a valuable checklist for practical measurement issues to consider in doing quantitative research.

Related Topics

Vocabulary assessment, vocabulary tests, vocabulary learning, research designs, research methodology, validation

References


Measuring Vocabulary Learning Progress


