Overview

Psychometrics is, according to its first definition, “the art of imposing measurement and number upon operations of the mind” (Galton, 1879, p. 149). Psychometric assessment refers to the scientific measurement of psychological phenomena by linking observed test scores to hypothesized unobservable traits that are supposed to have caused or influenced the observed behavior. In this sense, these are indirect measures of relevant psychological phenomena. The present chapter is concerned with the measurement of psychological traits as independent variables in research on individual differences (IDs) in second language acquisition (SLA). The phenomena under investigation (e.g., motivation, working memory, or language aptitude) are referred to as latent traits, or constructs, and establishing their relationship with language learning requires high-quality tests that are able to detect individual variation.

The principal aim of the psychometric approach to test development in ID research is to produce measurement instruments that allow for valid interpretations of the individual differences revealed by the test scores. Test development in a broad sense also includes refinement or updating of already existing tests, but the focus of this chapter is initial test construction. The procedure involves three main parts. The first is to develop items to form a scale, the second is to assure that the scale yields consistent and precise measures (i.e., the scores are reliable), and the third is to verify that the test measures what it is supposed to measure (i.e., the test is valid for its purpose).

Individual differences in SLA have been investigated using psychometric tests developed both within and outside the L2 context. Examples of the first kind are tests of language aptitude (Carroll & Sapon, 1959), foreign language anxiety (Horwitz et al., 1986), or motivation for learning an L2 (e.g., Gardner, 1986). The second category includes tests of working memory, often developed for L1 reading research and later applied to SLA (e.g., Daneman & Carpenter, 1980), or tests of personality, perceptual style, or non-verbal intelligence, with their origins in different branches of psychology. Perhaps unsurprisingly, research using tests specifically developed in L2 research contexts has typically yielded the strongest relationships with language learning outcomes (c.f., Dörnyei & Ryan, 2015).

As evident from the various applications listed above, psychometrics is a truly interdisciplinary field of study. Although the main focus in this chapter is on psychological constructs that are thought to influence SLA, the theory and methods to be discussed are equally relevant to test development in many other areas where access to high-quality measurement instruments is para-
Psychometric Assessments

mount, such as in clinical or educational settings, or in the assessment of language—the outcome variable in SLA.

The chapter is organized in the following way. First, the main steps in test development are outlined and then item analysis and scoring are discussed focusing on the two main approaches to test construction: classical test theory (CTT) and item response theory (IRT). Then, methods for evaluating scores for reliability are introduced, followed by different approaches to validity. The final sections of the chapter will examine how psychometric methods have been reflected in SLA research on IDs, and how to improve psychometric awareness in future ID research.

Technical Features

This section gives an overview of the main issues in the construction, refinement, and validation of psychometric tests. It includes a general outline of the test development process, item analysis, test scoring, and procedures for evaluating the reliability and validity of test scores. To start with, psychometric test development is characterized by an ordered procedure with careful attention to the function of individual items. High-quality items are the building blocks that allow a test to produce reliable and valid scores. To this end, the procedure for test development consists of a series of steps, including:

(i) Defining the purpose of the test and the construct(s) involved
(ii) Writing items following detailed specifications
(iii) Pre-testing and piloting of items
(iv) Item analysis and revision
(v) Field testing on a sample representative of the intended test population
(vi) Examination of score reliability and validity evidence

Steps (i) and (ii) concern content sampling and the choice of item format. Unlike achievement tests, instruments measuring cognitive ID traits typically do not have well-defined content domains (Irwing & Hughes, 2018). Instead, the test content is defined based on a description of the kinds of behavior that are likely to tap the construct of interest (e.g., remembering a list of spoken nonwords in a test of phonological memory). Expert judgments and existing research and theory may guide content selection (e.g., previous tests aimed at the same construct). The initial steps in test development also involve decisions about item format, which may depend on what is being measured. A useful distinction is between tests of optimal (or maximum) performance and tests of typical performance (Crocker & Algina, 2008). In optimal performance tests (in which more of the trait is “better”, e.g., in cognitive aptitude tests), items often have one single correct answer presented together with one or more incorrect alternatives (e.g., multiple-choice items). Tests of typical performance (in which more of a trait is “different” but not always “better”) often employ Likert-type items, in which respondents indicate on an ordered response scale to what extent they agree with a statement (e.g. “I often feel nervous in the language classroom”). Some tasks in ID research may require that test takers supply a response (e.g., repeating digits or words in a test of working memory) and not simply selecting between given alternatives. Despite differences in formats, the basic principles for item analysis are similar.

Item Analysis

Steps (iii) and (iv) above involve performing item analysis, which is at the heart of the psychometric approach to test construction and evaluation. It is a fundamental procedure to ensure the reliability and validity of test scores. Item analysis includes computing item statistics, among which the item difficulty and the discriminating power of an item are the most central. In classical test theory
(CTT), the item difficulty of dichotomously scored items is defined as the \( p \)-value (not to be confused with the probability value in tests of statistical significance), which refers to the proportion of test takers who responded correctly to a particular item. The discriminating power of an item is the most important technical feature in ID testing because it allows the separation of individuals along the measurement scale. Separating individuals by their level of measured ability increases the total test variance, which is critical for producing reliable scores that may correlate with other variables. For dichotomously scored items, difficulty and discrimination are related, because item variance is defined as \( p(1-p) \), and is thus largest when the \( p \)-value is 0.50. (Crocker & Algina, 2008).

With multiple-choice items, one also needs to allow for guessing, meaning that somewhat higher \( p \)-values are preferable (depending on the number of response options). Large item variances in turn contribute to total score variance, which is essential for discriminating between test performances (Crocker & Algina, 2008). From this perspective also follows that items with a \( p \)-value of 0 or 1 (i.e., none or all of the test takers passed the item) are in fact useless for the measurement of individual differences, since they do not contribute to variance. In attitudinal or motivational Likert-type items, the concept of difficulty often makes no sense. However, in item analysis, the mean response to a Likert item is equivalent to a \( p \)-value, and larger variance increases the discriminating power of the item.

Item discrimination may be expressed as an index of the strength of association between an item and the total score of the test. There are different ways to compute a discrimination index but one common method is the point-biserial correlation between the individual item and the total test score, with the item itself excluded (Crocker & Algina, 2008). Items that discriminate well (i.e., they work in the same direction as the rest of the test) have higher item-total correlations, meaning that they separate respondents with high overall scores from respondents with low overall scores. A point-biserial correlation above 0.20 in a sample of 100 participants (and slightly higher in smaller samples) is a reasonable benchmark (Crocker & Algina, 2008). Items with low or negative total score correlations should be revised or discarded from the test, since they impair reliable measurement. Examining item-total correlations is often an efficient method to detect items that are confusing to test takers for some reason, or items that have been mistakenly keyed in the wrong direction.

An alternative approach to item analysis is to use item response theory (IRT; for accessible introductions, see DeMars, 2010; Hambleton et al., 1991). In this framework, item difficulty and discrimination are estimated as parameters of a function that simultaneously estimate the latent ability of the test taker. The function finds the probability of a correct response to an item, given the ability of the examinee and the estimated item parameters. Person ability and item difficulty are expressed in the same measurement unit, which allows the researcher to evaluate item difficulty in relation to each examinee and not just on average for the whole group of respondents, as in classical test theory. A useful feature of IRT models is the visual representation of item functioning as item characteristic curves, produced by most IRT software. The most well-known IRT models for dichotomously scored items are the one-parameter logistic (1PL) model, which only describes items in terms of difficulty; the 2PL model, which also estimates item discrimination; and the 3PL model, which estimates an additional parameter for the probability of making a correct guess (the latter is useful in multiple-choice tests). IRT models are also available for polytomously scored items, for which the probability of each response category is estimated, given the examinee’s standing on the latent trait (DeMars, 2010; Hambleton et al., 1991). A limitation for small-scale research projects is that samples need to be large for reliable item parameter estimation with the 2PL and 3PL IRT models (DeMars, 2010). However, the dichotomous (1PL) Rasch model has been used in research also with smaller samples (see Bond & Fox, 2015 for an accessible introduction to Rasch measurement). Rasch analysis places particular emphasis on using item fit statistics, making it a practical tool for item analysis. Items that are misfitting a Rasch model are often those that would have poor discrimination in a CTT analysis, but the
Psychometric Assessments

Rasch analysis may provide additional information about at what ability level the item is not functioning properly.

Yet another approach to item analysis has been developed within the common factor model (McDonald, 1999). With this model, factor analysis of the item variance-covariance matrix produces estimates of item loadings onto a common factor representing the latent trait targeted by the test. The item-factor loadings (i.e., correlations between item scores and the latent factor) correspond to discriminatory power in classical test theory; higher item-factor loadings allow for more accurate measurement of the latent trait.

In addition to analyzing item difficulty and discrimination, it is often useful to examine the behavior of all response options in tests that employ a multiple-choice item format. The wrong options, called distractors, should appear attractive to an examinee who does not know the correct answer, and the keyed option should appear unambiguously correct to someone who knows the answer. A distractor may be defective for various reasons. It may appear highly implausible and thus be eliminated at once by respondents. This leaves fewer alternatives to choose from, resulting in a greater number of correct guesses by low-ability test takers. A distractor may be accidentally keyed as the correct response or it may actually provide an acceptable alternative that the test constructor initially failed to realize. Situations like these would generally result in poor item discrimination or item-model misfit, in which case a distractor analysis may reveal why an item does not work as expected. Methods for performing distractor analysis include the inspection of frequency tables with \( p \)-values for each response option, comparing response patterns of more- and less-able percentile groups (Haladyna, 2015), or techniques originating in item response theory (Thissen et al., 1989).

Test items may sometimes yield different item characteristics (e.g., \( p \)-values) for test takers with the same level of the measured ability but who belong to different groups (e.g., gender or ethnicity). This phenomenon is known as differential item functioning (DIF) and implies that group membership influences the item score over and above the measured ability. Importantly, this does not refer to situations when there are actual ability differences between the groups. Techniques for detecting DIF include the Mantel–Haenszel chi-square procedure (Osterlind & Everson, 2020) for comparing item success rates in two groups that are similar with respect to the measured ability. For Rasch fitted items, Bond and Fox (2015) suggested plotting the item difficulties obtained from two groups against each other and look for items that do not fall on a diagonal line (i.e., they deviate in difficulty in one of the groups). Items that display DIF should be inspected and, if motivated by the intended future test use, revised or discarded from the test.

**Scoring Issues**

The procedure that combines item scores into a total test score is often straightforward—individual item scores are simply added together—but there are situations in which test scoring needs more consideration. This includes when scores are corrected for guessing, or when items are nested within trials, or when the total score refers to a latent trait. These situations will be briefly addressed next.

Most assessment instruments in ID research use selected-response-item formats, in which guessing is always a possibility for test takers who do not know the correct answer. The fewer the response options, the higher the expected score from blind guessing. To deal with this problem, formula scoring (Nunnally & Bernstein, 1994) is sometimes used—for example, subtracting wrong answers from right answers while also considering the number of response options and avoiding negative scores. Critique against formula scoring has pointed out that guessing is seldom completely random, and that it is guided by test-taking strategies and partial knowledge (Lord et al., 2008). Another method, mainly applicable to large samples, is to fit a three-parameter logistic IRT model to the data, thus taking guessing into account when estimating the ability of the respondents.
A special case when guessing may be expected when binary choice items require respondents to differentiate between known and unknown stimuli (e.g., yes/no tasks in recognition memory research). In such tasks, signal detection theory (Macmillan & Creelman, 2005) is well suited to account for both hit rates (i.e., percentage of correctly identified known stimuli) and false alarm rates (i.e., percentage of unfamiliar stimuli incorrectly reported as being known). High false alarm rates are commonly encountered when people display certain response biases (e.g., always selecting “yes” when uncertain). Response bias may be compensated for by computing the total score by subtracting the standardized false alarm rate from the standardized hit rate (producing the $d$ measure; see Macmillan & Creelman, 2005).

When tasks consist of sets of items that in turn contain discrete, dichotomously scored elements, the computation of a total score may need some additional consideration. Some working memory span tasks are constructed in this way, and the choice of scoring method may then influence reliability or even the rank between participants. For example, respondents may be required to remember elements (e.g., words) belonging to items of increasing size (i.e., increasing number of elements within each item). One scoring method is to define memory span as the largest item size mastered by the respondent to some preset degree, for example four out of five correctly recalled items (absolute memory span). Another method is to count all correctly recalled elements in the entire set of items (partial credit scoring). Conway et al. (2005) provided detailed guidelines for scoring working memory span tests. Importantly, computing partial credit scores was recommended in order to retain as much variance as possible, thus increasing total score reliability.

A third issue that may influence scoring decisions is the relationship between observed (manifest) scores and constructs. Often, the sum of observed item scores is of less interest than the hypothesized latent trait supposed to have influenced them—the former is an indicator of the latter and a higher sum score does not necessarily correspond to a higher latent score. For example, in a two-parameter IRT model, more discriminating items are given more weight in the ability estimation. Latent trait estimation includes IRT modeling and factor analysis (both briefly introduced above concerning their role in item analysis). These methods generally require large samples that may often be out of reach for small-scale research projects, but they present some clear advantages. IRT models always produce latent ability estimates and, under certain conditions (a well-fitting Rasch model), the trait is measured at interval level, meaning that scores on the measurement scale represent equidistant latent trait levels. Factor analysis is normally used to investigate relationships among variables and factors, but may additionally be used to estimate a factor score for each respondent on each factor when there is more than one variable targeting approximately the same construct. Thus, by combining different tests tapping the same construct, it may be possible to better represent the trait in question in subsequent analyses (e.g., in correlations with SLA outcome variables). Generally, IRT and factor analytic approaches are compatible but specific research questions may guide the choice of model (e.g., a greater focus on individual items would warrant an IRT model).

**Reliability**

Having arrived at an observed total score, one wants to know if the same score would be obtained if the test was repeated under similar conditions (assuming that the individual’s ability has not changed). This is the problem of reliability, which constitutes a major topic in psychometric research. It will be treated here mainly in a CTT context and then briefly in relation to IRT.

In classical test theory, an individual's observed test score is conceptualized as a random variable consisting of two parts: the test taker's true score and an error component. The true score is a respondent's expected score over repeated test administrations. Reliability is defined as the ratio of true score variance to observed score variance or, alternatively, as the correlation between perfectly parallel tests (Lord et al., 2008). The part of the observed variance that is not attributable to
the true score is error variance. Error in the true score model refers to random error that changes non-predictively between test administrations, with an expected value over repeated measures that equals zero.

Reliability can be estimated in two ways: using different test administrations or using a single administration. Different administrations of the same test to the same participants allow for computing the test–retest correlation between scores over a time interval (coefficient of stability) or between parallel test forms on the same occasion (coefficient of equivalence). From a single test administration, which is most often the case in research, coefficients of internal consistency may be computed. The most widely used in SLA research is coefficient alpha (Cronbach, 1951), which estimates reliability as a function of the ratio of summed item covariances to total score variance, and provides a lower bound point estimate of the population reliability. It is appropriate to report with unidimensional measures (meaning that the items target one single construct or factor), but a high alpha is, however, not evidence of unidimensionality (Miller, 1995). Dimensionality in the test can be evaluated with factor analysis, after which coefficient alpha may be reported for each set of items that are associated with a particular factor (Cortina, 1993). It is generally misleading to report a single internal consistency coefficient for a multidimensional instrument, for example a test battery consisting of several subtests targeting different constructs.

Critique against the ubiquitous use of coefficient alpha has been raised and alternatives proposed (Dunn et al., 2014; Sijtsma, 2009). One common objection is that alpha comes with strict assumptions of the dataset that are seldom met in research, in which case the coefficient loses its precision. An alternative to alpha with less strict assumptions is coefficient omega (McDonald, 1999), which is now included in many software packages (cf., Dunn et al., 2014 for an accessible guide on how to use this coefficient). Whereas coefficient alpha assumes equal loadings for all items on the common latent factor of the scale (i.e., equal discrimination), coefficient omega does not have this restriction and may thus be particularly useful when item discriminations differ substantially.

Reliability will be affected by factors that change the ratio of true to error variance in a set of scores. One such factor is group heterogeneity. More between-subject variance in the measured ability means that the true score component increases (assuming that error is random and group independent), thus increasing reliability. Another factor is test length—adding more items (similar in difficulty and content) increases the amount of true score variance relative to error, following the Spearman–Brown formula (cf., Crocker & Algina, 2008, p. 145). Thus, adding new items is an important complement to improving existing items if the researcher wants to achieve higher score reliability. Correspondingly, factors that increase the error variance will have a negative effect on reliability, such as disturbances during test administration (e.g., noisy environment and unclear instructions), unengaged respondents, and deficiencies in test items discussed earlier.

When reporting reliability, the following may be kept in mind. First, because reliability is defined in terms of variances obtained from a particular test administration, it is sample dependent and not an inherent quality of a test. Researchers reporting reliability coefficients from other studies or test publishers should be aware of this and explain in what context the adopted coefficients were actually computed (or preferably compute a new coefficient based on the actual sample). Second, reliability coefficients reported in research are often accompanied by a value judgment as to if the reliability is enough, good, or even high. Nunnally and Bernstein (1994, p. 265) noted that this depends on the purpose of the study and suggested that for high-stakes psychological assessments, a reliability coefficient of 0.90 is a minimum, whereas a “modest reliability” of 0.70 could suffice in research where important decisions about individuals are not the aim. The evaluation of internal consistency coefficients should also consider test length, because the longer a test is, the lower inter-item correlations are needed to produce a high value of alpha (Cortina, 1993). When reporting individual scores, the reliability coefficient is used for computing the standard error of measurement to form confidence intervals around scores (Crocker & Algina, 2008). The standard
error of measurement in CTT constitutes an average of all test takers’ errors and is the same for all scores, independent of their position in the score distribution.

Item response theory models (introduced earlier) yield estimates of reliability comparable to that of the CTT model but with more flexibility in estimating variation in measurement error at different levels of the ability scale. IRT conceptualizes reliability in terms of item information, such that an item is most informative when its difficulty is near the ability of the examinee; its discriminatory power is high and the impact of guessing is low (Hambleton et al., 1991). The level of item information is expressed in an item information function that produces different values depending on the ability of the examinee. Adding item information functions from a test together produces the test information function, which corresponds to the reliability coefficient in CTT. Because the value of the function varies along the ability distribution, it provides different information (or “reliability”) for individuals depending on their ability. The level of test information is related to measurement precision, which may be expressed as a standard error of estimation (analogous to the standard error of measurement in CTT) and used to create a confidence interval about the individual test score. Standard errors tend to become larger at the extremes of the score distribution and smaller in the center (DeMars, 2010). The Rasch method can also produce separate reliability coefficients for persons and items (test scores may have more precision in measuring one or the other), and reliability for person abilities can be expressed as a separation index of how many distinct levels of ability the instrument reliably can distinguish (Bond & Fox, 2015). In sum, IRT provides individualized measurement information, whereas CTT yields reliability estimates common to all examinees. The advantages of IRT, however, come at the cost of the larger sample sizes needed to produce a good model fit to the data.

Validity

Validity concerns to what extent a test measures what it is hypothesized to measure. Validity has traditionally been conceptualized as content-related, criterion-related, and construct-related validity (Nunnally & Bernstein, 1994). More recent developments in validity theory have promoted a unitary view of validity, subsuming many kinds of validity evidence in a coherent interpretive argument (Kane, 2006). This subsection will discuss validity both from the traditional view and unitary view, in that order.

Content validity refers to whether item contents are relevant to the task and representative of their domain. In tests of individual differences in SLA, theoretical considerations, interviews, observation, or teachers’ descriptions of more- and less-successful learners may serve as sources of input for content decisions. Content validity is often examined using expert judgments and in many situations, problems with content validity can be detected by a thorough item analysis in which items with odd parameter values may be flagged for closer inspection. Two important concepts in validity theory are particularly relevant to content selection: construct under-representation and construct irrelevant variance (Messick, 1989). The first refers to situations when the test does not cover the attribute it is intended to measure (e.g., if a test of motivation in L2 learning only contains items about classroom learning but is used to measure motivation in other contexts). The second concept refers to situations when irrelevant factors influence the test scores in a way that was not intended by the test constructor (e.g., if the same test of motivation is written in English but used with speakers of other languages with low levels of English who may not understand the questions).

Criterion validity refers to the association between a set of test scores and some external criterion (e.g., scores from another test). It is often divided into concurrent validity (when the test and criterion scores are obtained simultaneously) and predictive validity (when the criterion measure is obtained at some later time, e.g., after a language course). Concurrent validity is of interest when developing a new test form of an existing test, to investigate if they measure the same thing. Predictive validity is of particular importance in individual differences studies because we often
want tests to predict SLA outcomes. Criterion validity is typically evaluated with correlational methods to obtain an effect size of the strength of association between the test and its criterion, and the most common measure to this end is the Pearson correlation coefficient. When the scope is to evaluate the prediction of future performance by more than one independent variable, regression analysis is preferred. Regression is, in many respects, similar to correlation but the aim is to identify the joint and unique contributions of each of the predictor variables to the outcome variable. (cf., Crocker & Algina, 2008 for a non-technical explanation of these methods). Several problematic issues surrounding correlational research are closely related to low reliability. Importantly, this pertains to both the measurement of predictor variables (i.e., individual differences) and outcome variables (i.e., the language tests used to measure learning outcomes). One issue concerns the small sample sizes often encountered in SLA research. Although correlations may turn out to be statistically significant, confidence intervals are often large, suggesting measurement imprecision. What constitutes a “small” sample depends on the purpose of the study. To achieve highly accurate predictive validity in a population, samples in the hundreds are necessary (Crocker & Algina, 2008), but for small-scale research projects, samples of 30–50 are sufficient for many purposes. Finally, two well-known problems in correlation studies are the restriction of sample range (e.g., due to floor or ceiling effects), resulting in low correlations, and the dichotomization of variables (e.g., by median split, upon which group means are compared). Dichotomization in ID research effectively discards information about individuals by lumping them together in groups and is generally not recommended because it mostly underestimates, but sometimes causes, spurious results (MacCallum et al., 2002; Plonsky & Oswald, 2017).

Construct validity is concerned with finding evidence for hypothesized latent traits supposed to influence test performance. Methods to evaluate construct validity include correlational techniques but unlike how these are employed in the evaluation of criterion validity, the focus is on explanation rather than prediction. Campbell and Fiske (1959) pointed out the importance of examining both convergent evidence (i.e., positive associations with similar constructs) and discriminant evidence (i.e., weak or no associations with dissimilar constructs) in order to establish construct validity. They suggested examining this with a multitrait–multimethod technique, whereby different constructs are tested with different methods. Evidence supporting construct validity would be if measures of the same construct using different methods would have higher positive correlations than measures of different constructs using the same method. A related, but computationally more intricate, method commonly employed to establish construct validity is confirmatory factor analysis. This method is used to examine to what extent a set of observations fit a model of predetermined relationships between variables. Using this technique, the researcher may compare different hypothesized relationships that are supposed to exist between latent variables. A comprehensive but accessible introduction to confirmatory factor analysis is found in Kline (2016). Finally, construct validity may also be examined with qualitative methods that are intended to tap into the mental processes active in test takers while engaged in a task. An example is verbal reports (Ericsson & Simon, 1993) during which participants verbalize their thoughts while answering items. The researcher then analyzes the obtained think-aloud protocols in order to detect if expected cognitive processes were involved in solving the task.

The unitary model of validity (Kane, 2006) subsumes the traditional forms of validity along with considerations of both test internal (scoring and reliability) and test external features (e.g., decisions based on test scores). An advantage of this model over the traditional view of validity is that it highlights that different sources of validity evidence are related and not compartmentalized as different validities. The model represents validity as a series of inferences that together justify interpretations and uses of test scores. Validity is supported to the extent that empirical and logical evidence makes each inference reasonable. Kane (2006) suggested that the inferences to be considered are scoring, generalization, extrapolation, and implication. Due to limited space, they will only be explained briefly and somewhat simplified here. The scoring inference from
observed performance to a test score is supported if items are correctly scored and with sufficient fit to a scaling model. If the task involves raters, their coherence is evaluated at this stage. The generalization inference from an observed score to a universe score has a function similar to reliability analysis in classical test theory. The term “universe score” refers to an individual’s expected score over repeated administrations of similar items (the universe of generalization). Conditions supporting this inference are equivalent to those supporting reliability (e.g., the sample of observations is large enough to limit random error). The extrapolation inference from a universe score to a target score refers to the representativeness of the scores in the larger context (the target domain) that the test scores are supposed to be informative about. An example would be to what extent a test of spelling can represent writing ability in general. Methods for supporting this inference are similar to those used to examine content and criterion validity described earlier. Finally, the evaluation of implications concerns the extent to which test scores provide evidence to support trait interpretations, including theoretical consequences (e.g., evidence supports a theory) or practical consequences (e.g., a student is admitted to a program based on test scores). Methods for evaluating the implication inference are similar to those used to examine construct validity.

In sum, the traditional and the unitary approaches to validity are similar as to the kinds of evidence they draw on, but the latter is more inclusive and relies on a more coherent conceptualization of validity.

Contributions to ID Research

This section will highlight a few examples of how the techniques outlined above have been used in the development and evaluation of psychometric tests of language aptitude and working memory. Conative and affective individual differences (e.g., in motivation or anxiety) are mostly investigated with survey methods, which are treated in Chapter 25 (this volume). Likewise, the measurement of implicit learning is addressed in Chapter 30 (this volume). The discussion will follow the subheadings in the previous section, in the hope that this will foreground the methods rather than individual tests or studies.

At the item level, only a few studies have featured detailed reporting from the test development process. Li & Luo (2019) used Rasch item analysis in the development of a language aptitude test, utilizing the visual features of Rasch software to graph item difficulties and person abilities on the same scale in a very readable manner. The study is a good example of transparency in test development at the item level. Rasch analysis was also used, together with classical item analysis, in a study by Bokander & Bylund (2020) on the internal validity of the LLAMA language aptitude tests (Meara, 2005). They observed that many items produced low parameter values, which they suspected could explain the low reliabilities often found with this aptitude battery. The authors suggested that a revision of the LLAMA tests would be beneficial to the L2 aptitude research community.

At the test score level, reliability estimates make important contributions to ID research because they provide information about the level of measurement error in a study. Reporting reliability coefficients may also guide researchers when deciding on which test to use. In working memory studies, the reliability of span tasks has been extensively researched in cognitive psychology, consistently producing high coefficients (Conway et al., 2005). Research using the modern language aptitude test (MLAT) has sometimes reported reliability coefficients from the original study (Carroll & Sapon, 1959) in which they were very high. This does not, however, guarantee that they will be so in every new study using this test. Reliability estimates from LLAMA aptitude test scores were reported by, for example, Bokander and Bylund (2020) and Granena (2013). Both these studies found that the reliability of some subtests was low, meaning that findings based on these tests should be interpreted with care. If an ID test gains widespread use in the research community before its
Psychometric Assessments

Psychometric properties have been established, reliability issues may have negative consequences for the field’s understanding of the investigated trait.

Validity evidence from correlation coefficients with criterion measures is an important source of knowledge about how ID variables influence or are associated with language acquisition. Language aptitude tests have generally produced high correlations with SLA outcomes compared to other ID variables, explaining some 10–30% of the variance in criterion scores (Li, 2016), and most of this research has used the MLAT (Carroll & Sapon, 1959). The MLAT has been translated into several languages but, as noted by Li (2019), translations have been significantly less predictive of language learning outcomes than the original version. A means of increasing criterion correlations was demonstrated in Conway et al. (2005). They included different working memory span measures and used factor analysis to produce factor scores that reflected the latent trait common to the different tests. The factor scores yielded higher correlations with the outcome measure than any separate memory span task.

Construct validity evidence is mostly obtained with correlational and factor analytic methods. Language aptitude has been examined in relation to other cognitive ID variables, for example working memory and implicit learning (Granena, 2013) and motivation (Gardner, 1986). Studies like these have served to define language aptitude constructs by producing convergent and discriminant validity evidence. A thorough demonstration of how to perform construct validation of an ID test is found in Grigorenko et al. (2000). The authors used factor analysis of a new aptitude measure and a large number of cognitive tests, which revealed a general component of cognitive ability and a language-specific component. Concurrent validity was evaluated with correlations with a more established aptitude test (MLAT) and predictive validity evidence consisted of correlations with teachers’ judgments about students’ language ability. Qualitative methods in construct validation have been rare in cognitive ID research but one example is Bokander & Bylund (2020) who used think-aloud protocols to examine respondent behavior in an aptitude task supposed to be implicit in nature. The verbal reports elicited from respondents revealed, however, that more explicit, conscious problem solving was involved than expected.

Finally, observing that validity evidence for the increasingly popular LLAMA language aptitude battery was either lacking or scattered about in different studies, Bokander & Bylund (2020) adapted the unitary validity framework outlined above (Kane, 2006) to propose a model for orderly aptitude test validation. They provided validity evidence at the inferential levels pertaining to items and test scores, and invited researchers to add evidence at the remaining inferential levels. A similar approach to validation would be possible to apply to any test of IDs in SLA.

Future Directions

A general impression from reviewing ID research in SLA is that studies rarely report psychometric details of how the test instruments were constructed, including information about item development and pilot testing. Too often, the reader simply has to trust that the instruments in the study produced reliable and valid scores. Future research should either include such information or direct the reader to sources where it can be evaluated and not simply be taken for granted. Without information about reliability and score distributions, it is hardly possible to ascertain whether a low or non-significant observed correlation with SLA outcomes was due to measurement error or actual lack of correspondence between the constructs involved.

Two future directions related to latent trait estimation merit some attention. First, a latent variable approach using multiple tests was briefly mentioned above and illustrated in Conway et al. (2005). By combining different but similar tests of a latent trait and performing correlations using factor scores instead of raw scores, more error-free measures and higher correlations were obtained. This seems to be an underexplored approach in ID studies in SLA, and future research may want to probe that possibility further. It is then necessary for the research community to develop many
tests targeting the same construct, preferably freely available as open source. Second, using item response theory methods, latent trait estimation may be more precise and tests can be administered that are tailored to the respondent's ability level (i.e., computerized adaptive testing). This would require an available item bank in which items are tagged with their individual measurement characteristics. As noted earlier in the chapter, such methods come with sample size requirements that are difficult to meet for small-scale projects, but like the multiple-test approach just described, this may be overcome by pooling resources. Important steps towards this aim have already been taken, with increasingly more publishers promoting open science (meaning that original datasets become available to researchers). The IRIS database (Marsden et al., 2016) is a platform where researchers can make their tests, and even raw data, openly available to the research community. This and similar initiatives mean that large amounts of test data could be obtained and analyzed, which would dramatically increase our prospects of developing high-quality measures of individual differences in SLA research.

References


Psychometric Assessments


