Linnaeus University Dissertations
No 422/2021

Lars Bokander

Validity considerations in the study of language learning aptitude

LINNAEUS UNIVERSITY PRESS

Validity considerations in the study of language learning aptitude

Linnaeus University Dissertations

No 422/2021

VALIDITY CONSIDERATIONS IN THE STUDY OF LANGUAGE LEARNING APTITUDE

LARS BOKANDER

LINNAEUS UNIVERSITY PRESS

Validity considerations in the study of language learning aptitude
Doctoral Dissertation, Department of Swedish, Linnaeus University, Växjö, 2021

ISBN: 978-91-89460-07-2 (print), 978-91-89460-08-9 (pdf)
Published by: Linnaeus University Press, 351 95 Växjö
Printed by: Holmbergs, 2021

Abstract

Bokander, Lars (2021). *Validity considerations in the study of language learning aptitude,* Linnaeus University Dissertations No 422/2021, ISBN: 978-91-89460-07-2 (print), 978-91-89460-08-9 (pdf).

Language learning aptitude is a hypothesized psychological construct that has been used to explain differences in how fast and how well people can acquire a second language (L2). It is generally assumed that language learning aptitude is a multidimensional phenomenon, meaning that it consists of sub-constructs that are not necessarily interrelated. Research on language aptitude and its relationship with language learning outcomes has been undertaken for at least 70 years but much still remains unknown about the nature of this construct. Key to understanding the effects of a hypothesized latent trait like language aptitude is to ensure that it can be meaningfully quantified, and also that whatever real world observations that the trait is supposed to be linked to (in this case, L2 acquisition) can be measured with sufficient accuracy. The present thesis set out to explore issues in the measurement of both language learning aptitude and its predicted outcome (L2 acquisition), specifically applied to a context in which the L2 is Swedish. The validity of an increasingly popular test of language aptitude, the LLAMA, was examined in detail and a test of Swedish receptive vocabulary for L2 learners (the SweLT) was developed with the aim of efficiently serving various research purposes, including the study of language aptitude effects. In addition, theoretical and methodological issues in the assessment of individual differences in second language acquisition were outlined. The results from the empirical studies suggest that the LLAMA suffers from imprecision but that it may still be useful in research if due care is given to the interpretation of the obtained test scores. For quick assessment of general proficiency in Swedish, the SweLT seems to be a promising candidate but further refinement of this test is called for. Finally, some possible implications of aptitude research are discussed, including future use of aptitude tests as practical tools for individual adaptation of educational programs for adult L2 learners of Swedish. The findings of this thesis make it clear that the LLAMA would not be suitable for this purpose.

Keywords

language aptitude; language testing; individual differences; Swedish vocabulary; test validation

Contents

List of original papers

Study 1: Bokander, L. (forthcoming). Psychometric assessment. To appear in S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge Handbook of Second Language Acquisition and Individual Differences*. Routledge.

Study 2: Bokander, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning 70*(1), 11–47.

Study 3: Bokander, L. (2020). Language aptitude and crosslinguistic influence in initial L2 learning. *Journal of the European Second Language Association, 4*(1), 35–44.

Study 4: Bokander, L. (2016). SweLT 1.0 – konstruktion och pilottest av ett nytt svenskt frekvensbaserat ordförrådstest. *Nordand*, 11(1), 9–30.

As to the division of labour in study 2, Lars Bokander designed and carried out the study, and was the principal author of the background, methods, results, and discussion sections. Emanuel Bylund authored most of the introduction and conclusion sections of the paper, as well as providing valuable feedback on the rest of the paper. Emanuel also contributed by recruiting several of the participants for the study. In study 1, 3, and 4 Lars Bokander was the single author (although with wonderful feedback from supervisors, reviewers, editors and others).

Acknowledgements

This thesis was written at the Department of Swedish at Linnaeus University in Växjö, Sweden. Throughout the years that I have spent here, I have always felt warm support from my supervisors (Manne, Nina, Gisela) and colleagues (too many to mention here). Thanks to all of you! I would also like to thank the editors and anonymous reviewers that have provided valuable, and sometimes harsh, feedback on my manuscripts. You have shown me that research is a truly collaborative undertaking. Sixhundred and forty individuals, most of them people that I have never met, kindly contributed data to this thesis. Without data there can be no research, so super-mega thanks to all of you, as well as to the persons who helped organize data collection on different sites around the world. I would also like to thank some people, or perhaps circumstances, that brought me here in the first place. For many years, before moving back to Sweden to pursue a PhD degree, I was working abroad in a non-academic setting with tasks involving recruiting, language teaching, and assessment. That period opened up my eyes to individual differences among language learners, as well as to methodological issues surrounding the measurement of L2-ability. Meanwhile I had the opportunity to pursue a masters degree in language testing by distance at Lancaster University. So big thanks to all my former colleagues at Samres in Chişinău, and to my exceptional teachers at Lancaster! Without you this thesis would never have existed. Finally, many acknowledgement texts like this one express something along the lines of 'thanks to my family for having endured my absence during these years, etc.'. Fortunately, it has not been like that for me at all. Being a PhD candidate allowed for plenty of freedom in planning my time, and I think that I have been rather present as a husband and a father. Ala and Mirela, I hope you agree, but if not, I love you anyway! The same goes for my parents, of course, to whom I dedicate this text.

Växjö
9 August 2021

1 Introduction

Human performance in most areas of life displays a great deal of variation between individuals. One such area is the learning of new languages in adulthood. Some individuals seem to pick up a second language (L2[1]) with ease and within a relatively short time they are able to attain a high level in the new language. For others it is not quite so simple and for a few, learning a new language may be an almost impossible task. Most second language teachers have observed these differences in their classrooms, particularly in more homogenous student groups that are otherwise similar with respect to, for example, first language (L1) and educational background. Second language researchers have hypothesized that differences in the ability to learn a language can, to some extent, be explained by a psychological trait called language learning aptitude (in this text also referred to as language aptitude, or simply, aptitude). Language learning aptitude has been defined as a special talent for learning languages, that (i) differs between individuals; that is (ii) relatively stable over time; (iii) that is different from general intelligence, and (iv) that is not influenced by previous language learning experiences (Skehan, 1998). These four points may of course be questioned, but they provide a good indication of how researchers have approached the concept of language learning aptitude. Language aptitude has been claimed to be the most influential individual difference variable in (adult) second language acquisition (Dörnyei & Ryan, 2015), yet much remains unclear about what it actually comprises and how accurately it can be measured. Different tests of language aptitude have been constructed and put to use in different contexts. It is not well known how much these tests have been used outside the domain of second language acquisition (SLA) research but they have been applied successfully in selection to language training programs by governmental agencies in the United States (Stansfield & Reed, 2019). It has often been pointed out by aptitude scholars that establishing language aptitude profiles for individuals may be helpful for making placement decisions in education, thus ensuring that each individual will receive language training that is tailored to his/her abilities and needs (e.g., Robinson, 2001).

In Sweden, the most obvious candidate for large scale placement decisions would arguably be introductory language programs aimed at adult immigrants (svenska för invandrare, SFI), comprising about 150,000 enrolled students as of 2019 (Skolverket, 2020). Today, group placements in SFI are normally based on immigrants' previous educational background in their home country.

[1] The division occasionally made in the literature between second and foreign language acquisition seems to me as a somewhat crude dichotomization of a complex variability between situations in which language learning takes place. In this thesis, L2 acquisition will refer to second language learning in general, and particularities relevant to the learning context will be specified when/if needed.

Available information about educational background may somtimes be scant and it is well known that there remains a large variation in language attainment between individuals in the adult language classroom even after this placement procedure (Skolinspektionen, 2018). Hypothetically, any variable that is known to covary with educational outcomes should be of interest when making group placement decisions. Besides educational background such variables may include, for example, typological distance between L1 and L2 (Ringbom, 2007), integrative or instrumental motivation (Dörnyei & Ryan, 2015), or language learning aptitude. Although language aptitude has (to the best of my knowledge) hitherto never been considered by Swedish policy makers as a selection tool for adult language training, it is not inconceivable that this could happen in the future. There has been some public dissatisfaction with the success of language programs in Swedish for adult immigrants (Skolinspektionen, 2018) and there are thus reasons to investigate the pros and cons of capitalizing on information other than learners' educational background when making placement decisions or designing language courses. Increased knowledge about the extent to which various background factors influence immigrants' success at acquiring Swedish as an L2 may be useful for informing organizational and policy decisions about how to most efficiently facilitate their L2 learning process. Potentially, such decisions may have far reaching consequences both for individuals and society. It is thus imperative that research findings informing those decisions are based on valid and reliable methodology.

In Sweden, L2 studies that include language aptitude as an explanatory variable have been scarce with a few notable exceptions (e.g., Abrahamsson & Hyltenstam, 2008; Agebjörn, 2021; Bylund et al., 2010). However, little attention was given to the validity of the instruments employed in those studies (e.g., about the aptitude measures employed, or to what extent a grammaticality judgement test can function as a proxy for L2 ability) and this can be said in general about much language aptitude research, where it is simply assumed that test instruments will perform as desired. In the field of second language acquisition (SLA [2]), the recent decade has seen increasing calls for a methodological reform in L2 research due to a growing awareness of the need to improve scientific rigor, allowing for more robust and replicable research findings (Gass et al., 2020; Gass & Plonsky, 2020; Marsden et al., 2016; Plonsky, 2013). This line of work has highlighted important concerns related to statistical analysis methods, statistical literacy in the research community, and research transparency (i.e., open science). Equally important in this quest is, however, an increased focus on test construction because an ever so sophisticated statistical method will only yield reliable findings to the extent that the data submitted to it were obtained with high quality measurement instruments.

[2] The acronym SLA will be used in this text both as referring to the academic discipline and to the process of acquiring a new language.

A central aim in this thesis is to outline key aspects of theoretical and methodological nature regarding test practices in research on language learning aptitude. This is done by examining validity issues in the measurement of both the independent (i.e., aptitude) and the dependent (i.e., language proficiency) variables that occur in aptitude research. The inclusion and analysis of an L2 proficiency measure in the thesis is motivated by the fact that one cannot discuss someone's aptitude for something without defining what that something is. Thus, the overarching research question guiding the thesis concerns the internal and external validity of a popular language aptitude test – the LLAMA (Meara, 2005) – and the internal and external validity of a vocabulary test developed by the author – the SweLT (Bokander, 2016). Study 1 (Bokander, forthcoming) outlines main methodological steps in test validation that were applied to varying degrees in Study 2–4. Study 2 and 3 address the internal and external validity of the LLAMA, respectively. Study 4 addresses the validity of a pilot version of the SweLT.

This introductory chapter is organized as follows. In section 2, I discuss relevant issues in validation and test theory. Then, in section 3 conceptualizations of language aptitude are reviewed and examples are provided on how language aptitude has been operationalized in tests. In section 4 I turn to the criterion variable – L2 proficiency – and I discuss to what extent receptive vocabulary may serve as a practical, and measurable, representative of overall L2 skills. Next, the individual studies are presented and their main findings are discussed in relation to the principal aim of the thesis and concerning possible future directions of language aptitude research.

2 Validation, test theoretical concepts and issues

This section introduces central terminology that is being used throughout the rest of the text and in the individual papers. It also brings up a few methodological issues that have involved some controversy among researchers, and defends the methodological choices made in the thesis.

2.1 A unitary view on validity

Inquiries about validity concern the extent to which a test or a research study provides valuable information for some purpose. Over the last century, validity in educational and psychological measurement has been conceptualised and defined in many different ways. This has given rise to an ample terminology, some of which I will briefly explain here to facilitate the reading of this introduction and the individual studies. Two main tendencies among scholars have been to conceptualize validity either as a unitary phenomenon, or as a fragmented set of different kinds of validity (Newton & Shaw, 2014). Recent texts on validity often contrast a 'classical' trinitarian view of validity that dominated psychological research in the 1950–70s (comprising *content*, *construct*, and *criterion* related validity; the latter often divided into *concurrent* and *predictive* validity depending on the purpose of the study) with attempts to describe validity as a unitary construct, mainly associated with the works of Messick (1989) and Kane (2006). In this thesis I will draw on Kane's model because it constitutes a convenient framework in which various kinds of validity evidence may be analyzed in a coherent way. Importantly, a unitary view on validity also means that methods for item analysis and reliability estimation (discussed below in this section) are both included as ways to establish the validity of test score interpretations. In earlier test theoretical approaches, such as the classical trinitarian view just mentioned, reliability was usually treated as being distinctly separate from validity concerns (Newton & Shaw, 2014). Bokander and Bylund (2020) proposed a structure for analyzing language aptitude tests with Kane's model, and the same structure can be applied to any language test. Although this was not explicitly done in Bokander (2016) a similar procedure for examining validity evidence was applied in that study, examining both test internal aspects (item functioning, reliability) and test external aspects (correlations with L2 proficiency). Terminology from the trinitarian view on validity (content, criterion, and construct validity) fits well into Kane's framework and will be used in this thesis as well.

Validation according to Kane (2006) follows a series of steps (called inferences) that, in test development, can be said to describe the entire

development process from test specifications and item trialing, to the evaluation of reliability, criterion and construct validity, and finally, test use for some pre-specified purpose, such as research or education. With Kane's terminology, these steps are the *scoring* inference; the *generalization* inference; the *extrapolation* inference; and *implications* of test use. The scoring inference, as interpreted in this thesis, concerns the item level of a test (item analysis and scoring). The generalization inference concerns issues related to reliability, and the extrapolation inference concerns criterion and construct validity, that is, relationships to other variables outside the test itself. Some interpretations of Kane's model have included an *explanation* inference to specifically address construct validity (e.g., Purpura et al., 2015) and this practice was adopted in the second study (Bokander & Bylund, 2020). Implications of test use are mostly outside the scope of this thesis but will briefly be addressed in the general discussion, Section 7 below. In addition, this introductory text as well as the individual studies use the terms *internal* and *external* validity. Test internal features include content, item properties and reliability, whereas external validity roughly corresponds to correlations with other variables (construct, or criterion validity). The terms are thus used here (and in much language testing literature) in a slightly different way as compared to experimental psychology where internal reliability concerns phenomena related to an experiment, whereas external validity concerns generalization to the 'real world' outside the laboratory (Shadish et al., 2002).

In general, the findings in this thesis are more robust for the internal validity of the two tests under scrutiny (LLAMA and SweLT), whereas external validity evidence plays a somewhat minor role. Internal test validation, that is, finding support for scoring and generalization inferences, was carried out with item and reliability analysis in study 2 (Bokander & Bylund, 2020) and study 4 (Bokander, 2016). These two methods, along with some issues surrounding them, will be discussed in what follows.

2.2 Item analysis

Item analysis is a central part of test development and plays an important role in two of the empirical studies included in this thesis (Bokander, 2016; Bokander & Bylund, 2020). The main reason for performing item analysis is to ensure that all items in a test contribute to measuring the intended construct, and that they do not introduce irrelevant variance in the test scores, for example, due to large measurement errors, or by tapping a different construct altogether. Two different methods for evaluating item functioning were employed – classical test theory (CTT), and item response theory (IRT) – more particularly, the dichotomous Rasch model (Rasch, 1960). The CTT approach to item analysis was described in detail in Bokander (forthcoming). It is very straightforward, mainly based on correlations and proportions of correctly answered items, so it

need not be further discussed here. The basic tenets of the Rasch model will be briefly explained here, following the presentation in Bond and Fox (2015). Unlike in CTT, Rasch modelling expresses person abilities and item difficulties in terms of *probabilities* of passing or failing an item. The probability of a correct response to an item is expressed as a function of the difference between a person's standing on the latent trait (i.e., the measured construct), and the difficulty of the item. The higher the person's ability, and the lower the item difficulty, the greater is the probability of a correct response. Unlike in much statistical modelling Rasch analysis does not attempt to describe the closest fit to observations, but it is more like stating a null hypothesis (an ideal measurement model) against which observations are compared. Real data never conforms perfectly to this model, and the analyst is interested in how much the data deviate from the Rasch model. Small deviations are tolerated, large deviations imply that the test does not produce accurate measurements. Better fit to the Rasch model means that information about an item or a test taker is more reliable. The individual response patterns (i.e., each test taker's response vector of ones and zeroes) in a dataset can be assigned specific probabilities, and the deviations between observed data and the Rasch model are reported as fit statistics (i.e., chi-squared tests of expected-to-observed score differences). Improbable response patterns indicate that either items, test takers, or both, deviate from the measurement model and are thus problematic from a psychometric point of view. Item difficulties and person abilities are measured on the same scale (in units called logits, i.e., the logarithm of the odds for a correct score on a particular item). The logit values for items and persons typically range from about -3 to +3 and for optimal measurement quality, one would like a test to include items that cover the entire range of test taker abilities. Often, this is not the case, and then floor or ceiling effects may contribute to lower reliability (which can be observed in study 2 and 4 of this thesis). Some critique has been directed against using Rasch measurement with items that allow for guessing, because guessing can inflate person ability estimation (Stewart, 2014; Stewart et al., 2017). However, because the aim in study 2 and 4 was to explore item functioning, rather than to accurately measure test takers' abilities, the issue with inflated person estimates was not considered a threat to the analysis.

2.3 Reliability

An important feature of the generalization inference in the validation framework (Bokander & Bylund, 2020) is about establishing reliability. Reliability is a necessary but not sufficient condition for detecting correlations between variables (i.e., supporting an extrapolation inference of a validation study). If a low correlation is found between two variables, but the test scores were

unreliable, it is not possible to know if the low correlation was due to an actual absence of association between the variables, or simply due to low reliability in one or both measures. The empirical articles in this thesis all used coefficient alpha as an estimator of reliability, mainly because alpha is the most widespread reliability coefficient in the behavioral sciences. However, the practice of expressing test score reliability with the coefficient alpha has come under increased scrutiny over the last decades, mainly in areas outside of SLA but recently also within language studies (Plonsky & Derrick, 2016). This section will highlight some of the main arguments that have been put up against the ubiquitous habit of reporting coefficient alpha as a reliability estimate.

Most criticism against using coefficient alpha (e.g., Dunn et al., 2014; McNeish, 2018) points out that alpha comes with strict assumptions about the data set, some of which are rarely met in reality. In particular, three such assumptions are commonly discussed. The first assumption is that of *unidimensionality*, which means that the test instrument, or the measurement scale for which reliability be estimated, targets only one single construct (i.e., a dimension). An example of noncompliance with this assumption would be a school assessment of, for example, Swedish history, which also requires a high degree of writing skills in order to score points for the questions. Such a test would involve both history and writing, and thus not be unidimensional. For the same reason it also makes little sense to report an internal consistency coefficient for a full test battery such as the LLAMA, in which different subtests are targeting different constructs. The second assumption of coefficient alpha, and probably the least respected one, is that of *essential tau-equivalence*. This means that all items on a test should have the same relationship to, or the same factor loading on, the measured construct (i.e., the common factor of a unidimensional test). This assumption can be tested with factor analytic techniques, but it is often enough to inspect the point biserial correlations between each item and the total score of the scale (i.e., the discrimination in classical item analysis). If these are markedly different, the assumption of essential tau-equivalence probably does not hold up. The third assumption is that of *uncorrelated errors*. In classical test theory, measurement errors are supposed to be random and thus uncorrelated. However, one situation when this assumption would most likely not be met would be a test that uses different item formats, for example, by mixing multiple choice and open-ended questions. In this situation, errors would be likely to vary with the different item formats and consequently, coefficient alpha would not be an appropriate estimate of the test score reliability.

The criticism against estimating reliability with coefficient alpha has also been countered. For example, Ryakov and Marcouliedes (2019) showed that for unidimensional measures with items contributing somewhat equally (but not necessarily tau-equivalent) to measurement, coefficient alpha does produce good reliability estimates. In essence, this amounts to following sound

psychometric procedures for test development (cf. Bokander, forthcoming), making sure that piloting and calibration of items in the end produces a test that meets the requirements of unidimensionality, approximately similar discrimination values, and are free from artefacts that introduce correlated errors. Sometimes, however, other principles than the strictly psychometric ones may guide test development, for instance content related priorities; one such case with relevance for the present thesis being frequency band based vocabulary tests. Although vocabulary size can be construed as a unidimensional variable, the inclusion of bands of markedly different difficulty are likely to violate the assumption of essential tau-equivalence. Thus, coefficient alpha is most likely a poor choice for reporting a single reliability estimate for such tests (see Schmitt et al. (2020) for a related discussion).

One of the most frequently proposed alternatives to alpha is coefficient omega (McDonald, 1999). It conceptualizes reliability similar to alpha but uses factor analysis to identify a common factor of a test instrument. Each item's loading on this factor is an indication of the contribution of that item to the whole test. If all items have the same loading on the common factor, they are tau-equivalent and the reliability estimate is identical to that of alpha. However, the computation of omega produces good estimates of reliability also when items load differently on the common factor, and this is one important argument that has been put forward in attempts to convince researchers to begin reporting omega instead of alpha (Deng & Chan, 2017; Dunn et al., 2014).

2.4 The point of reference in testing

Finally, an issue with implications for both item analysis and reliability is whether a test score is given a norm referenced or a criterion referenced interpretation. In norm referenced testing, a score is compared to other scores; in criterion referenced testing, a score is related to a criterion (e.g., mastery of 80% of the content in a language course). In norm referenced testing, the measured trait is assumed to be normally distributed in the population, and a good test should approximate that distribution as close as possible. This means that only a few test takers will have very high or very low scores, and that the scores will be spread out along the normal distribution, which in turn allows for maximal reliability due to the increased true score variance (cf. Bokander, forthcoming). With a criterion referenced test, on the other hand, most participants may theoretically pass the cut score and thus be considered to have mastered the test content. In this case, one is not concerned about variability among test takers that fall within either of the two possible scores (pass or fail), but a good test should maximize its sensitivity around the cut score (by selecting items at that particular difficulty level) in order to ensure reliability.

Research on individual differences in SLA or psychology is generally concerned with maximizing variability between individuals on the trait under

investigation, which assumes a norm referenced approach. That is also the theoretical view adopted in most of this thesis. However, frequency banded vocabulary tests (cf. Section 4 of this introductory chapter) have often been used with criterion referenced interpretations that seek to determine if a test taker masters a particular frequency level or not (e.g., Schmitt, Schmitt & Clapham, 2001). Mastery of a frequency level may in turn have pedagogical-practical consequences, for example in estimating how much vocabulary that can be covered in a text of a certain difficulty. Importantly, although methods for item analysis and reliability estimation may differ somewhat between norm referenced and criterion referenced testing (Brown & Hudson, 2002), it is the purpose of a test that determines what theoretical approach one takes in the analysis of test scores. Thus, the fact that frequency banded vocabulary tests have often been used for the purpose of establishing mastery/non-mastery of a certain frequency level, does not invalidate the use of a norm referenced approach with the same tests when they figure in individual differences research.

3 The predictor variable: language aptitude

This part of the thesis discusses the independent variable, that is, the construct of language aptitude; its different conceptualizations and attempts to measure it. The first subsection (3.1.) introduces some main lines of inquiry in aptitude research and the second subsection (3.2.) provides examples of the kind of test tasks that have been used in language aptitude research.

3.1 Aptitude theory

This section reviews some of the more influential models of L2 learning aptitude, including its possible relation to L1 development, and examines suggestions on how to complement traditional aptitude constructs with models of working memory and implicit learning ability.

3.1.1 The classic (Carrollian) model of language aptitude

Carroll (1981), in reviewing the history of language aptitude research, noted that early 20th century language aptitude tests were either targeting L1 ability (similar to the verbal components of an intelligence test), or were measures of achievement after a short introduction to the L2 to be learned. In many respects they were reflections of the grammar-translation teaching methods of that time (often in the teaching of Latin or ancient Greek), aiming more at fostering intellectual capabilities than to enable communication in a foreign language. These older aptitude tests were mostly inappropriate for predicting learning outcomes with the audio-lingual method developed by US army linguists. The latter usually employed trial courses in a foreign language as a selection tool. This was a successful approach for reducing dropout rates, because the trial course seemingly tapped into qualities necessary for learning an L2. However, it could not provide insights about which separate aspects of L2 learning were involved. It was also an expensive method, and hence, there were strong incentives to develop new, more valid aptitude tests. During the 1950s, this demand sparked the development of the Modern Language Aptitude Test (MLAT, Carroll & Sapon, 1959), as well as a new theory of language aptitude.

The Carrollian approach described language aptitude as comprising four sub-constructs – phonetic coding ability, grammatical sensitivity, rote-learning ability for foreign language materials, and inductive language learning ability (Carroll, 1962, 1981). Phonetic coding ability refers to the ability to identify distinct sounds and to retain associations with the symbols representing them.

Grammatical sensitivity is the ability to identify linguistic entities and their syntactic function in the sentence structure (without the use of grammatical meta-language). Rote-learning ability concerns the formation and retention of sound-meaning associations and inductive language learning ability refers to the ability to infer linguistic rules, given a sample of language (which contains the relevant structures to be discovered). Carroll arrived at these four dimensions by means of factor analysis of about 30 different tasks that had been hypothesized to predict foreign language learning during a one-week intensive trial course in Mandarin for L1 English speakers, all male recruits at the US Air Force. The tasks also included tests of interest in L2 learning, L1 verbal ability (e.g., verbal fluency, or associative memory) or tasks involving artificial languages.

The five subtests of the MLAT (for details, cf. section 3.2.1 below) were designed primarily for prediction and quick administration, and there was no intention to represent the four aptitude constructs in a balanced way (e.g., inductive learning ability was not represented in the MLAT). This may have contributed to the widespread opinion among SLA researchers that MLAT was atheoretical or an insufficient representation of language learning processes (Robinson, 2001; Skehan, 1998; Winke, 2013). One may also note that the exploratory factor analysis from which the four aptitude components were initially derived (Carroll, 1958) was done with a rather small sample of participants and Carroll himself admitted that interpreting the factor loadings was difficult. Skehan (1998) suggested that collapsing the constructs of grammatical sensitivity and inductive language learning ability into one single dimension of language analytic ability would make more sense, and this approach has since then been adopted by many (or most) aptitude researchers of today (see Li & Zhao, 2021). The L2 learning context that Carroll had in mind was situations in which the learner makes a deliberate effort to learn the language under formal instruction (Carroll, 1981, p. 83) and this has often been used as a point of criticism against the validity of language aptitude as conceptualized in the MLAT.

3.1.2 The aptitude-treatment interaction model

Observing that the MLAT was validated mainly with learners at the early stages of L2 learning, and in learning contexts dominated by classroom application of the audiolingual L2 learning method, Robinson (2001, 2005) argued that aptitude theory and aptitude tests should be extended to encompass longer time intervals and other contexts of L2 learning, for example naturalistic, untutored language acquisition in an L2 environment. Robinson's theory conceptualizes aptitude as complexes of cognitive abilities that vary between individuals and are differentially predictive of success for different kinds of L2 tasks. Tailoring aptitude tests to L2 tasks was seen as particularly important for learners with highly differentiated ability complexes, who may benefit from some particular

teaching method but fail under another. Because a primary interest was directed towards the long term relationship between aptitude tasks and real life L2 performance, Robinson suggested that aptitude tests should also target pragmatic and interactional abilities – features that were not considered in traditional aptitude testing.

Robinson's (2005) interactive aptitude model combined core cognitive abilities not uniquely related to language, aptitude complexes and tasks in initial input-based learning, aptitudes for advanced language tasks, and real life L2 use with complex demands on pragmatic and interactional skills. An example of how these different aptitude combinations interact, is the well known situation of teachers providing feedback in the form of recasts. Initially, processing speed and pattern recognition (a complex of basic cognitive abilities) enable the learner to 'notice the gap', whereas phonological working memory capacity and speed (another basic ability complex) enables memory for contingent speech. Noticing the gap and memory for contingent speech in turn make learning from recasts possible, and the efficiency of this particular instructional method is determined by individual differences in the basic cognitive aptitudes described. The aptitude-treatment interaction approach advocated by Robinson has continued to attract attention in experimental research on aptitude effects in relation to classroom feedback, instruction types, practice strategies, and task complexity (Li & Zhao, 2021).

3.1.3 Aptitude and processing stages in SLA

Addressing the lack of theoretical rationale behind the MLAT, Skehan (1998, 2002, 2016) proposed to connect the study of language aptitude to acquisitional stages in SLA, suggesting that different aptitude sub-constructs be differentially important at different stages. He hypothesised several areas of SLA where individual differences were likely to be found, some of them already targeted by existing aptitude tests and others not. Among several areas that had not yet been included in aptitude tests were most notably working memory and its language related sub-components (cf. below the subsection on working memory as language aptitude).

The staged model states, with some variation between different versions, that SLA begins with an input processing stage during which segmentation of the incoming linguistic signal takes place. This initial stage is highly dependent upon phonological memory and attention control, which are considered to be important functions of working memory (Baddeley, 2003). More phonological memory allows for longer stretches of language to be processed. Then, follows the noticing stage in which the learner discovers the links in the language between form and meaning. One of the traditional aptitude components, phonetic coding ability, seems particularly important to create such form-meaning links. Also, individual differences in phonological memory are hypothesized to play an important role, similar to the initial input stage, because

the more information that can be held simultaneously in the phonological loop, and the longer it may be stored there, the higher is the probability that the information can be incorporated into subsequent processing and storage in long-term memory. Next, the pattern identification and pattern restructuring stages involve language analytic ability (including grammatical sensitivity and inductive language learning ability). Executive working memory combines information from long-term memory and new linguistic input, and updates L2 knowledge as new information is added. Executive working memory also governs attentional control, which is important for explicit form-focused language learning. Especially in instructed SLA settings, attentional control is conducive to error avoidance and incorporation of feedback into performance. Individual differences in executive working memory and language analytic ability are thus strong determinants for how fast a learner develops language structure beyond the most basic features of the L2. The final parts of the staged model, the pattern control stages, concern long term SLA and these are the less well described and somewhat more speculative parts of the model (Skehan, 2019). They include automatisation and lexicalisation, which refer to both fluency and vocabulary development. One important process at this stage is chunking of language material, that is, the creation of longer stretches of language that are quickly available to the speaker without having to be consciously analysed. Chunking is a necessary process for the development of fluency and long-term memory representations.

Skehan (2019) observed that most aptitude research in recent decades has focused on the earlier acquisitional stages, those that relate to handling sound and patterns in the L2. The later stages, involving automatisation and proceduralization, have to some extent been included in the HiLAB which is the most recent major aptitude test battery (Linck et al., 2013) but much work remains to be done. One challenge involved here is that examining later stages of development require longitudinal designs that may stretch over decades.

3.1.4 Working memory as language aptitude

Individual differences between language learners in working memory capacity has attracted a growing interest among L2-aptitude researchers in recent decades (Wen et al., 2017). The concept of working memory has been developed in several models in cognitive psychology (Miyake & Shah, 1999) but the most influential model in aptitude related SLA research has been the multicomponent model of working memory first described by Baddeley and Hitch (1974, see also Baddeley, 2003). The model specified a central executive function that controls attention and processes material that is temporarily held in either of two subcomponents for storage and rehearsal of information – the phonological loop (for auditive information) and the visuo-spatial sketchpad (for visual information). A later version of the model (Baddeley, 2000) added a feature called the episodic buffer, to account for short term storage of combined

information from different sensory modalities and from long term memory, into a unitary representation.

Most research on WM in SLA has concerned the executive function and the phonological loop. The phonological loop was described by Baddeley as a verbal memory device that holds a small amount of information while repeating it continuously until it fades away after a short amount of time (temporal decay). The performance of the phonological loop is thus dependent on both its storage capacity limit (i.e., how much information that can be held at one time) and for how long time the information can be rehearsed/updated. As long as information is kept in the phonological loop, it is available for processing by the central executive which controls attention and performs operations on the temporarily stored material. Cognitive tasks differ to the extent to which they involve the central executive function. For example, repeating a series of digits that one has just heard does not involve any complex information processing and it is thus not assumed to depend on the central executive. However, repeating the same digits backwards implies that one holds the digits in the phonological memory while also performing the operation of reversing their order. This is a more complex task that involves the central executive to perform the necessary operations while the digits are kept accessible in the phonological loop. Often, the term working memory is reserved for the latter kind of executive functioning, whereas the term phonological short-term memory (PSTM) is used to denote the simpler form of memory without additional information processing (Cowan, 2008). These two parts of the working memory have shown differential relationships with SLA processes. The PSTM has, for example, been reported to predict (at least initial) vocabulary development (Gathercole, 2006; Service & Kohonen, 1995; Speciale et al., 2004) whereas the complex (executive) WM appears more related to language analytic ability and general, long-term, L2 development (Juffs & Harrington, 2011; Linck et al., 2014). Traditional aptitude test batteries have shown a modest relationship with working memory tests. However, measures of both simple and complex working memory were included in the Hi-LAB aptitude test battery (Linck et al., 2013).

3.1.5 Aptitude and L1 ability

The most well known longitudinal research on language aptitude is arguably a series of studies by Sparks and colleagues (Sparks & Ganschow, 1993; Sparks et al., 2009), who followed children from preschool age to high-school, at which point they completed the MLAT battery and their grades in foreign languages were obtained. The researchers were thus in a position to consider language aptitude in relation to the participants' L1 development and general intelligence, which had been documented at different points during the study. Particular attention was directed towards phonological coding, which was defined by Sparks and Ganschow (1993) as the "ability to sequence, break down and put

together the sounds of language" (p. 297). Importantly, the word *phonological* should not be primarily associated with pronunciation skills. The following quote explains the term phonological coding in detail. "Although it may have relevance for pronunciation, the term specifically refers to the ability to discriminate between speech sounds, learn sound/symbol correspondences, and identify sound segments (phonemes) within words" (Sparks and Ganschow, 1993, p. 297).

Sparks and colleagues carried out several studies with US high-school students of foreign languages (mainly German, French or Spanish as an L2) in which they early on observed that students with strong L2 skills and high scores on the MLAT test (Carroll & Sapon, 1959) had usually strong L1 skills as well, and conversely, students weak in L2 and low scores on the MLAT usually performed below average on L1 tests (Sparks & Ganschow, 1993). Further retrospective studies involving archived records from elementary school (i.e., age 6–9 years) revealed that L1 performance in the early school years were highly predictive of L2 achievement and L2 aptitude in high-school (i.e., when the students were about 16-18 years old). This finding was corroborated in longitudinal research in which Sparks and colleagues followed children over ten years in school (Sparks, 2012; Sparks et al., 2009).

Taken together, these studies showed that early L1 skills, in particular word decoding, spelling and vocabulary, was highly predictive of future performance on the MLAT and L2 language courses. Word decoding and spelling depend on phonological processing and Sparks and Ganschow (1991) proposed the Linguistic Coding Deficit Hypothesis (LCDH) which stated that weak L1 skills in childhood, for example, due to insufficient phonological processing, were strongly related to difficulties in learning subsequent languages. As pointed out by its originators, the LCDH is reminiscent of Cummins's (1979) hypotheses of linguistic interdependence (i.e., L1 and L2 learning draws on the same underlying abilities) and threshold effect (i.e., the level of L1 moderates subsequent L2 learning). Being special educators of children with learning and reading disabilities, Sparks and Ganschow naturally focused on the less able L2 learners, so it is not clear if the LCDH extends to the other end of the learner distribution (high-performers). Phonological processing ability as defined by Sparks was, however, suggested to be normally distributed in the population, meaning that there is no categorical or substantial difference between children diagnosed with language disorder and those who are just weak, but subclinical, language learners. Support for this also comes from large twin studies on L2 learning in the field of behavioral genetics, suggesting that L2 learning ability is continuous and normally distributed, and also highly heritable – even more so than L1 development (Dale et al, 2012). Interestingly, the studies by Sparks and colleagues convincingly demonstrated that the MLAT seems to tap into abilities that are necessary for successful L1 development, indicating that valid language aptitude tests should consider participants' L1 skills. This poses a

potential problem for language aptitude research that I will return to in the general discussion.

3.1.6 Aptitude for more or less conscious learning

It is well known that adult SLA includes, at least in the beginning, very conscious attempts to learn the L2 (e.g., memorizing new vocabulary, or grammar rules). However, some L2 learning also seems to proceed unconsciously, as observed in early SLA as a distinction between learning and acquisition (Krashen, 1985). With a more up-to-date terminology, the two phenomena are known as explicit and implicit language learning. Implicit learning has been defined as a process resulting in knowledge that is not fully accessible to consciousness and that is difficult to verbalize. It concerns the learning of relatively complex and abstract information, and learning takes place incidentally and without awareness, even though attention to the task is required (Seger, 1994). A challenge has been to establish that the learning and the resulting knowledge are indeed unaware to the respondent. Methods to investigate awareness include retrospective verbal reports and confidence ratings (Rebuschat, 2013).

The study of implicit learning is closely related to research on statistical learning (Conway et al., 2010; Perruchet & Pacton, 2006). To various extent both the implicit and the statistical learning traditions seek to explain subconscious language processing that is dependent on sensitivity to patterns, frequency information and transitional probabilities between language constituents, in the development of fluent and automatic language use. The statistical learning tradition has often involved young children and infants (Saffran et al., 1996) whereas implicit learning studies have been done with adults (in which explicit metacognition is developed). Sometimes the terms are used without distinction, and the notion 'implicit statistical learning' has been proposed in the literature to cover both (Christiansen, 2019). They are used with much the same meaning throughout this thesis.

Language aptitude research has recently suggested that traditional aptitude tests, like the MLAT, primarily appear to be tests of explicit learning ability and that aptitude studies should include tests aimed at implicit learning as well (e.g., Granena, 2013b, 2019; Linck et al., 2013). The Hi-LAB included two subtests of implicit cognitive processes because of a potential association between implicit learning ability and the development of long-term, high level language skills. The serial reaction time task (SRT, described below in section 3.2.6) successfully differentiated between learner groups at different L2 proficiency levels. Granena (2013a) suggested that LLAMA D may be tapping implicit learning aptitude because it displayed weak relationships with explicit tasks, but a closer association with a test of implicit learning. This finding, however, did not replicate well in another study (Granena, 2019), in which LLAMA D

displayed a weak relationship to other implicit tasks, despite a large sample of participants. Others have found evidence for explicit, conscious strategies being employed when test takers complete LLAMA D (Bokander & Bylund, 2020; Suzuki, 2021) meaning that findings based on this test could not be taken as evidence of implicit learning. The evidence is, thus, still inconclusive as to the impact of implicit language learning aptitude on L2 development. This stems at least in part from the unreliability of implicit tasks in general, and the relatively small number of aptitude related studies that have addressed the issue.

3.2 Testing aptitude

This section will examine the content of four aptitude test batteries that have made contributions to the study of language aptitude, each in their own way. In addition, examples of working memory tasks and an implicit learning task are described because they represent two areas that have shown promise as language aptitude components in recent studies (Wen et al., 2017). Although only LLAMA was used in the present thesis, a brief survey of other related tests may give the reader a general impression of the kinds of tasks that have been included in language aptitude research. The presentation also serves as a backdrop against which the LLAMA tests will be discussed later. Other aptitude test batteries than the ones described here have been published and used in research (e.g., Parry et al., 1990; Petersen & Al-Haik, 1976; Pimsleur, 1966) but the examples given below cover most of the kinds of tasks that have been used in aptitude testing. Because the acronyms of the tests are so well established, they appear as rubrics in following subsections.

3.2.1 MLAT

The MLAT (Modern Language Aptitude Test) was developed by Carrol and Sapon (1959) and is still arguably the most used aptitude test battery in language research. A detailed account of the work that led to the publication of the MLAT can be found in Carroll (1962). The test comprises the following five parts. In part I, Number Learning, test takers have to learn the number system of an unknown language, presented auditorily with English translation. In a subsequent test phase, they are prompted to write down (with digits) the numbers that are read aloud to them. In part II, Phonetic Script, test takers learn phonetic notations for some common English sounds and are then tested by selecting one out of four alternative spellings to a spoken word. Part III, Spelling Clues, is a speeded task in which test takers are presented with English words that are written with a spelling that approximates their pronunciation (e.g., ‘kao’ for ‘cow’). They are then asked to select a synonym to each stimulus word, out of five alternatives. In part IV, Words in Sentences, each item consists of two sentences. In the first sentence, one word is underlined. The test taker then

indicates (out of five alternatives) which word in the second sentence has the same function as the underlined word in the first sentence. In part V, Paired Associates, test takers memorize, during two minutes, written foreign vocabulary with English translations. They are then presented with the foreign words and prompted to select the correct translation, out of five alternatives. The MLAT has become the most widely used language aptitude test and has been translated to several languages, including French, Japanese, Hungarian and even Braille (Stansfield & Reed, 2019).

Two observations, with relevance for the general discussion in section 7, can be made from this brief description of the MLAT. First, all five tasks make use of the test takers' L1, English. This means that MLAT needs to be translated to other languages if used with non-English L1-speakers. Second, the MLAT subtests are relatively long (the number of items in each task are 43, 30, 50, 45, and 24, respectively) and the multiple-choice format (in Part II to V) uses five response options. Both test length and many response options are features that tend to increase the reliability of test scores by reducing measurement error (other things being equal). This, in turn, allows for maximizing correlations with other variables, and indeed, the MLAT has consistently produced among the highest correlations in language aptitude research (Stansfield & Reed, 2019).

3.2.2 CANAL–FT

A rather different approach to language aptitude is represented by the theory CANAL-F (Cognitive Ability for Novelty in Acquisition of Language – Foreign), operationalised in the test battery CANAL-FT (Grigorenko et al., 2000). The CANAL-F theory has its base in cognitive psychology and postulates that language learning requires the ability to handle ambiguity and novelty, similar to learning other daily life tasks. The test is dynamic, meaning that measurement takes place over different occasions, and it uses an artificial language, *Ursulu,* which is gradually learned by the test takers throughout the test. The test battery consists of nine parts, five of which are administered on a first occasion and four (similar, corresponding tasks) at a later point in time (after at least 30 minutes), aiming to obtain measures of delayed recall from long-term memory. In comparison with the MLAT, the tasks in CANAL-FT bear more similarities to real language learning and stimuli are presented both visually and orally. In the first part, test takers infer the meaning of unknown Ursulu words interspersed in an English text. Part two is similar, but involves comprehension of whole Ursulu passages and not just individual words. The third part of CANAL_FT is a paired associates task but unlike MLAT part V, and more like real language learning, the words are semantically related to each other to facilitate retention. In part four, sets of sentences are presented to the test takers, who are then prompted to infer the meaning of a new sentence that is presented to them. Finally, in part five, participants are given short sentences

in Ursulu and are required to work out some simple grammar rules and vocabulary. This part bears some similarity to LLAMA F (cf., 3.2.3 below). In sum, CANAL-FT puts much emphasis on inductive language learning, which is an aptitude dimension that was hypothesised by Carroll but not represented in the MLAT. The CANAL-FT was validated in a carefully designed study (Grigorenko et al., 2000) which included the MLAT, tests of non-verbal and verbal intelligence, a prior language experience questionnaire, as well as language teachers' judgements of the participants' L2 ability. The correlation with L2 achievement was about the same as MLAT, and factor analysis revealed two dimensions; an intelligence related and a language-specific factor. Although Grigorenko et al. (2000) reported high reliability coefficients for all subtasks, and performance was similar to the MLAT in predicting language learning success, the CANAL-FT has never gained much attention among researchers and has appeared in only a few published language aptitude studies.

3.2.3 LLAMA

The LLAMA (Language Learning Aptitude Master of Arts program) was developed by Paul Meara and his students (Meara, 2005). The test suite, consisting of four subtests, was inspired by the Carrollian theory of aptitude (cf., 3.1.1) but it features two important innovations that have arguably contributed much to the popularity of this test battery. First, they are computer administered and free to download from the Internet. Second, they are based on picture stimuli and use words and phrases from indigenous American languages that are supposedly unknown to most prospective test users. This feature has led to the tests being perceived as 'language independent' (unlike the MLAT and CANAL-FT, described above) although Latin characters are used, which at least hypothetically should disadvantage test takers who are unfamiliar with latin script.

The LLAMA test suite consists of subtests B, D, E and F, and a common design is that each subtest has a practice, or stimulus, phase which is followed by a test phase. In LLAMA B, *vocabulary learning*, test takers have two minutes to learn 20 word-image pairings. In the test phase they are presented with words and are prompted to click on the corresponding image on the screen. In LLAMA D, *sound recognition*, the computer plays ten spoken phrases to which the test takers should listen carefully. In the test phase, 30 phrases are played by the computer and the test taker has to indicate which ones are new and which ones were present in the stimulus phase. In LLAMA E, *sound-symbol associations*, the test taker has two minutes to learn mappings between a set of symbols (consisting of Latin letters, digits and diacritics) and one-syllable sounds that are played (e.g., *pi* or *ma*) when a symbol is clicked by the participant. In the test phase, 20 two-syllable 'words' are played by the computer (e.g., *mapi*) and the test taker has to decide which out of two alternative spellings that

corresponds to the two-syllable word. In LLAMA F, *grammatical inferencing*, the test taker views pictures of figures; their shapes, numbers, colors and spatial relations. Each picture is described by a sentence in an unknown language. In the test phase, twenty pictures are displayed on the screen and the test taker must decide which of two alternative sentences that correctly describe the picture.

The LLAMA subtests mainly seem to target the aptitude constructs proposed by Carroll (1962) – paired associates, sound-symbol associations and grammatical sensitivity (and possibly inductive ability). However, LLAMA D seems to be a unique and innovative addition to the family of language aptitude measures (cf. Bokander & Bylund, 2020; Granena, 2013). In comparison with the MLAT, it may be observed that the LLAMA subtests are relatively short and (except subtest B) utilise a two-options response format. Both these features (other things being equal) are likely to contribute to the relatively low reliability found in studies using the LLAMA.

3.2.4 Hi-LAB

The High-level Language Aptitude Battery, Hi-LAB (Linck et al., 2013) differs in several ways from earlier language aptitude tests. It is based on SLA theory and contains 13 cognitive and perceptual measures of which many do not have counterparts in other aptitude test batteries, for example tests of implicit learning ability and working memory. The aim of creating the Hi-LAB was to predict long-term, ultimate attainment in natural learning contexts, which sets it apart from the MLAT-tradition of predicting early L2 acquisition in a mostly formal learning environment. Space does not permit a detailed review of all the Hi-LAB tasks here but some observations can be made. The tasks included in the test battery aim to operationalize the following set of constructs, with the number of tasks for each construct given in parentheses. Executive working memory (4); phonological short-term memory (3); associative memory (1); long-term memory retrieval (1); implicit learning (1); processing speed (1); and auditory perceptual acuity (2) (Linck et al., 2013:535). From this list it may readily be noticed that the researchers behind the Hi-LAB accommodated many of the language aptitude constructs proposed in the post-Carrollian period, most notably different aspects of working memory (e.g., Skehan, 1998; Wen et al., 2017).

Validation studies with the Hi-LAB are still scant, but some comparatively strong associations between Hi-LAB performance and high level L2 learning outcomes have been reported and include, for example, high classification accuracy into different levels of the ILR scale used in assessment at the Foreign Service Institute (Linck et al., 2013). However, because one principal aim of this test battery was to predict long-term learning, the research community still awaits results from longitudinal reports. In addition, being the product of a US Government financed project, and with an administration time of about 2.5

hours (Doughty, 2013), the Hi-LAB has hitherto not been easily accessible for the aptitude research community.

3.2.5 Working memory tasks

Linck et al. (2014) provided a comprehensive meta-analysis of working memory tasks in SLA, of which I will mention the ones most frequently used. Traditionally, most working memory tasks that have been employed in SLA research are memory span tasks in which the respondent is asked to repeat some recently presented information. A distinction is commonly made between tasks that aim to tap executive functions, and those aiming to test short-term memory span (the phonological loop) without involving the central executive (cf., section 3.1.4 above). These are referred to as complex and simple working memory tasks, respectively. A simple task that is assumed to not involve executive functions is the nonword repetition task, in which respondents are simply asked to repeat nonwords. Real words are avoided because they allow for mnemotechnical strategies and language knowledge to contaminate the task. Complex tasks, on the other hand, typically require the respondent to perform two simultaneous operations, one processing operation and one storing operation. The backward digit span task, mentioned earlier, is an example of such a task because the respondent must keep the digits in active memory storage while also reversing their order before repeating them. Another complex working memory task involving both storage and processing is the reading span task (Daneman & Carpenter, 1980) in which respondents read sentences (typically, two to six) after which they are prompted to recall the final word of each sentence. The reading span task has been modified to involve other types of processing stimuli, such as arithmetic operations instead of sentences (Turner & Engle, 1989). Operation span tasks are well suited for SLA research because there is less confounding with language skills than in a reading span task.

From a psychometric perspective it may be noted that all of the WM tasks mentioned are supplied response tasks, meaning that measurement error is typically much smaller than in selected response tasks that allow respondents to guess. WM tasks therefore tend to produce highly reliable data, in comparison with some of the aptitude tasks discussed above, or the implicit learning tasks to which we turn next.

3.2.6 Implicit learning tasks

Arguably, the implicit learning task that has been most represented in language aptitude research is the serial reaction time (SRT) task (Nissen & Bullemer, 1987). The respondent sits before a computer screen on which a stimulus can appear in any of four different locations. As soon as the stimulus appears, the respondent presses a corresponding button as fast as possible. In some experimental conditions, the stimuli appear in random order and in others, the stimuli appear in a structured sequence. Alternatively, stimuli may consist of

less probable or more probable sequences (Kaufman et al., 2010), creating a probabilistic SRT task. Crucially, respondents are not informed beforehand about the different sequences that may appear. Over trials, respondents become more accustomed to the non-random or more probable sequences which results in shorter response latencies (reaction times), indicating that implicit learning has taken place.

Like many other tests of implicit learning, the SRT task tends to produce scores with low internal consistency (e.g., Cronbach's alpha for the SRT data was .44 in the study by Kaufman et al., which the authors claimed as normal for that kind of task; see also Buffington et al., 2021 for a related discussion of reliability and validity of procedural memory tasks in SLA). Low reliability in a set of test scores generally limits the possibility of finding correlations with other variables and makes the test unsuitable for detecting individual differences between respondents (Bokander, forthcoming). It has even been suggested that the familiar dissociation between implicit and explicit learning (e.g., the former process being supposedly intact in amnesic patients, whereas the latter is not) may to some extent be a methodological artefact due to the low reliability associated with implicit learning tasks (Buchner & Wippich, 2000). Instead these two kinds learning may be located on one and the same dimension of consciousness (see Hulstijn, 2015, for a related discussion). Thus, research on individual differences in implicit statistical learning ability faces a significant challenge in developing tests of increased reliability. Suggestions on how to improve these tasks were made in Siegelman et al. (2017) and include, for example, increasing the number of items, using items with varying difficulty, and minimizing measurement noise by adapting test difficulty so that the participants perform well above chance levels.

3.3 Section summary

Language aptitude has been conceptualized and researched in a number of different ways and with different scopes. The Carrolian era prioritized, somewhat atheoretically, predictive validity for selection purposes, whereas later research approached language aptitude more from a cognitive perspective, trying to find out what language aptitude consists of, and its relation to SLA. Recent decades have seen attempts to bring in additional aptitude constructs, such as working memory and aptitude for implicit learning. The four representative aptitude test batteries described in this section seem to present different strengths and weaknesses. The MLAT has demonstrated high predictive validity but it has been criticized for weak theoretical underpinnings. The CANAL-FT was thoroughly founded in cognitive theory and has demonstrated reasonable construct and criterion validity, but it seems to presuppose a high degree of test language (i.e., English) literacy. The LLAMA is easily accessible to researchers and quick to administer, but there might be

some doubts surrounding the validity of scores reported from this test. The Hi-LAB seems to be able to outperform previous tests as far as construct and criterion validity is concerned, but its administration is time consuming and it has not been readily available to researchers. Working memory tasks are reliable and relatively simple to construct; implicit learning tasks are somewhat more intricate in design and tend to produce less reliable scores.

4 The outcome variable: ability in the L2

This part of the thesis turns to what language aptitude tests are supposed to predict, that is, language proficiency. It will be argued that measures of vocabulary size are good indicators of general language proficiency, and different ways of conceptualizing and quantifying vocabulary knowledge will be reviewed. The section closes with a discussion of possible aptitudes for vocabulary development. The fundamental idea behind considering vocabulary in aptitude research is that (i) vocabulary constitutes a central part of ability in an L2, and (ii) a good aptitude test should be able to predict L2 development, so (iii) a good aptitude test should be sensitive to L2 vocabulary acquisition.

4.1 Vocabulary and general L2 proficiency

Before examining the role of vocabulary in overall L2 ability it is necessary to define what it means to know an L2, that is, defining the construct of L2 ability. This is clearly not an easy thing to do and there have been many different proposals to describe L2 ability and how it may be assessed. Because this part of the introductory chapter aims at evaluating and to defend the use of vocabulary tests in language aptitude research, I will limit the discussion to L2 construct definitions made in the language testing literature. Accounts of what constitutes L2 ability in the modern history of language testing often begin with a "skills-and-elements" model of language knowledge (Lado, 1961, cited in Purpura, 2004. p.51). This model described three dimensions, or elements, of language knowledge (phonology, structure, and lexicon) that would be expressed, receptively or productively, in either of the four macro skills (listening, reading, speaking, and writing). This resulted in a three-by-four matrix aimed at assisting the creation of language tests, by specifying targets for discrete point tasks (e.g., written structure, or spoken vocabulary). Others have proposed similar models, including more fine grained description of its elements, but the basic idea of a matrix remains. The skills-and-elements model reveals a focus on content validity; the aim seems to have been to make sure to cover all (formal) aspects of second language learning. A different approach, guided by a focus on construct validity, was that of Oller (1979) which, using factor analysis, tried to find a single higher order factor that would explain general L2 ability – much like the *g* factor in models of human intelligence. This language factor would be best assessed by integrative tests tapping into many skills simultaneously and inspired by Gestalt psychology, Oller proposed the cloze test (originally devised for L1 reading diagnostics) as a useful tool for tapping integrative L2 knowledge. However, a number of factor analytic studies

have since suggested that a multifactorial model provides a better description of L2 ability, and most researchers, including Oller himself, eventually abandoned the idea of a single factor model for L2 ability (Bachman, 1990).

Following the 1970s' growing interest in communicative language teaching, a model of communicative competence for teaching and assessment purposes was laid out in a seminal paper by Canale and Swain (1980). This model was greatly influenced by Hymes's (1972) proposal to include both language knowledge and language use in a model of L1 communicative competence, thus disputing Chomsky's (1965) view that the study of language competence should not be concerned with language performance. Canale and Swain (1980) explicated communicative competence as consisting of two main parts – grammatical competence, and sociolinguistic competence – as well as metacognitive strategies to cope with situations in which the speaker's grammatical or sociolinguistic competence is not sufficiently developed. Grammatical competence in Canale & Swain's model included phonology, syntax, lexicon, and semantics, that is, most of what constitutes form and meaning in a language. Sociolinguistic competence, in contrast, concerned language use in real life situations, for instance, how to express politeness or other pragmatic aspects of using an L2. Bachman (1990) suggested a reorganization of Canale and Swain's (1980) model that has arguably become the most influential model of L2 ability in the field of language testing. Similar to Canale and Swain, Bachman juxtaposed aspects of language knowledge (grammar, phonology, lexis, discourse) which he called organizational competence, with aspects concerning language in use (pragmatic and sociolinguistic skills) which he called pragmatic competence. Bachman also developed the concept of strategic competence (as compared to Canale & Swain, 1980), which he claimed was involved in all language use and not just in situations in which the speaker lacks some necessary language ability to perform in the L2. There is much more detail in the model by Bachman (1990) that has been omitted here, but the interested reader is referred to the original text or to Bachman and Palmer (1996) which presented essentially the same model with some features relabeled.

The models of L2 ability so far discussed all gave a prominent role to vocabulary and this is even the case with the single factor model by Oller (1979), evident by his suggestions to assess language with cloze tests, a method in which lexical items are to be supplied by the test taker. What is not obvious in these models is how the different components are related to each other and to what extent performance in some language activities is dependent upon command over other aspects in the model. It also sometimes seems difficult to delimit what is more central to language ability from aspects that are less central. Hulstijn (2011, 2015a) suggested a division between core and peripheral language proficiency, thus making explicit how to weight the various parts of a communicative competence model. In core linguistic cognition Hulstijn

included the phonetic-phonological, morphosyntactic and lexical domains, as well as the speed or fluency with which these core features may be accessed, whereas strategic and metacognitive competences were seen as more peripheral components. His arguments for assigning different weights to different parts of language ability was that, first, more peripheral language proficiency cannot exist without its core components, whereas the inverse relationship does not apply. Second, factor analytic studies have usually revealed a first factor dominated by the core language components, and third, regression analyses have found that the core proficiencies explain a large part of the variance in language performance outcomes (Hulstijn, 2011).

It is thus mainly an empirical question to what extent one aspect of communicative L2 ability, or a subset of aspects, may represent overall L2 proficiency. High correlations between facets of a communicative model would imply that testing one type of knowledge will yield information about the others, whereas low correlations would imply that many different abilities need to be included in a test in order to tap L2 ability as a whole. Because an important issue in this thesis concerns to what extent L2 vocabulary may validly be used as criterion in language aptitude research, it is of interest to examine correlational evidence of relationships between L2 vocabulary and general L2 ability. There are several reasons for wanting to represent overall language ability with a narrower construct. One is that language testing is time consuming, and developing broad measures of language proficiency is resource intensive. In situations when research participants' scores from well known and validated tests of general language proficiency are available, researchers may prefer to use them (e.g., the TOFEL, if the target L2 is English). Arguably, no such test of Swedish exists today. Without access to general L2 ability tests, researchers must opt for the second best, that is, a test that is cheaper to produce but that would correlate highly with a general ability test. A well designed vocabulary test seems excellent for this purpose.

The close relationship between lexical knowledge and overall proficiency in an L2 is very well documented. Hulstijn et al. (2012) investigated the relationship between rated communicative L2 proficiency and linguistic correlates. Based on communicative adequacy in a set of speaking tasks, L2 learners were graded as level B1 or B2 according to the CEFR scale. They were also tested on a range of non-communicative linguistic tasks, out of which in particular mid and low frequency vocabulary knowledge was an efficient predictor of rater assigned CEFR level. In de Jong et al (2012), speaking proficiency ratings were regressed on nine hypothesized predictors of speaking performance. Vocabulary and ratings of pronunciation (intonation) explained about as much variance in the outcome variable as all the linguistic measures taken together. The role of vocabulary in speaking was further supported in Uchihara & Clenton (2018) Both vocabulary size and depth measures have been shown to predict performance in reading (McLean et al., 2020; D. Qian, 1999;

D. D. Qian, 2002), listening (Noreillie et al, 2018; Staehr, 2009), writing (Crossley et al, 2012; Schoonen et al, 2011;) and general L2 proficiency (Zareva et al., 2005). The LexTALE (Lemhöfer & Broersma, 2012), a lexical decision task developed for quick administration in psychological research, was found to be clearly superior to self ratings of L2 proficiency. Lexical knowledge has even been demonstrated to be more important for communication than grammar knowledge (Qian & Lin, 2019). For example, a structural equation model in Zhang (2012) indicated that vocabulary size was a better predictor of reading comprehension than a grammaticality judgement task. Hulstijn (2015) points out, however, that it may not be meaningful to compare lexis and grammar because they tend to develop hand in hand, and from a usage-based theoretical perspective, lexis and grammar may not be possible to separate.

This review of studies on vocabulary and L2 performance clearly shows that vocabulary is highly correlated with many communicative language skills. In language aptitude research, however, grammar has attracted most attention, whereas vocabulary tests have been rare. Li's (2016) metastudy investigating the construct validity of language aptitude, included 66 primary studies of which only seven included a vocabulary test as L2 criterion. The examples in the previous paragraph suggest that using vocabulary assessments in language aptitude research seems highly warranted, but it is as yet an underexplored area. One may then wonder what kinds of vocabulary tests would be best suited for this purpose, which is a question that requires a closer look at L2 vocabulary as a construct. I will briefly review relevant aspects of how L2 vocabulary has been modelled (i.e., what constitutes vocabulary knowledge) and operationalised (i.e., how vocabulary is measured) in previous research. To limit the subject, I will only be concerned with elicited and discrete vocabulary measures that by their construction may be said to belong in a psychometric assessment tradition, whereas vocabulary research based on free production (e.g., measures of lexical diversity) will be excluded.

4.2 Conceptualizing and testing vocabulary

Vocabulary in L2 development has been conceptualized as a multidimensional phenomenon. Different researchers have similarly proposed (under different labels) to conceptualize vocabulary ability in three dimensions, related to breadth, depth, and fluency (e.g., Daller et al., 2007; Henriksen, 1999). The first dimension, breadth, refers to how many lexical items a person knows. The second, depth, refers to how much is known about each word. The third, fluency, refers to the speed of access to words in the mental lexicon during fluent language use (see Gyllstad, 2013). The first and second dimensions have been frequently targeted in vocabulary assessment for research and educational purposes (Schmitt, 2014). The third dimension seems to have been somewhat less observed; perhaps due to the more technically involved procedures of

measuring time variables (but see Zhang & Lu, 2014). The second dimension is obviously important because it concerns how to use vocabulary, how to inflect words, how to combine them with other words, or knowing when a word may or may not be socially acceptable (Nation, 2013). However, as often pointed out (e.g., Milton, 2009; Schmitt, 2014; Vermeer, 2001), the depth dimension is vaguely defined, and has mostly been measured with instruments of questionable reliability and validity. This thesis will thus be concerned with the first dimension, vocabulary size, because that dimension seems more promising for reliable measurement (Gyllstad, 2013), which is crucial in research related to individual differences in other traits, such as language aptitude.

To some degree, operationalizing vocabulary size in assessment instruments must still involve depth aspects, because different item formats may tap word knowledge of different strengths. Laufer and Goldstein (2004) proposed a two-by-two matrix for constructing vocabulary items based on (i) the lexical information to be supplied by the respondent (word meaning, or form), as well as (ii) how that information is retrieved (by recognition, or recall). This creates four possible item types that have been frequently employed in vocabulary testing, and Laufer and Goldstein showed that the four item types form an implicational hierarchy of difficulty, such that (from difficult to easy) form recall > meaning recall > meaning recognition > form recognition. The first two essentially correspond to translation from L1 to L2, and from L2 to L1, respectively, although form recall has also been elicited by means of a context sentence in the L2 (Laufer & Nation, 1999). The third item type requires respondents to choose between different alternatives when presented with a word definition (e.g., Nation & Beglar, 2007) or a context sentence (e.g., Bokander, 2016). Finally, an example of the fourth item type is the Yes/No checklist test format (e.g., Lemhöfer & Broersma, 2012; Meara & Buxton, 1987) in which respondents state if they are familiar with a presented word, or not. Receptive (i.e., recognition) vocabulary is generally larger than productive (i.e., recall) vocabulary but they seem to be predictable from each other. Webb (2008) found that productive and receptive vocabulary sizes were more similar at higher frequency bands and when partial word knowledge was considered. In lower frequency bands and with full knowledge required, the gap increased between receptive and productive vocabulary.

When designing the SweLT, one purpose was to come up with a test that could be used with large samples of speakers with different L1s, and it should thus preferably be automatically scored and not involve translation to L1 (i.e., meaning recall). Criticism has been raised against the widespread use of vocabulary recognition test formats (as in SweLT), because they allow for guessing which can inflate vocabulary size estimates (Stoeckel et al., 2020). It has also been pointed out that most vocabulary size tests are imprecise due to low sampling rates of words per frequency band (Gyllstad et al., 2015). However, neither of these points seem to be a reason for abandoning meaning

recognition based tests like the SweLT, in situations where the aim is merely to reliably separate participants (as in individual differences research), rather than establishing absolute vocabulary size, or command of some particular frequency band. They do imply a problem, though, for the validity of SweLT if the purpose were to estimate vocabulary size (investigated in the fourth research question of study 4, this thesis).

A final point concerns test content, or more specifically, how to select vocabulary items. Selecting words to represent different frequency bands means that the test constructor prioritizes content validity. Representative content sampling then becomes a highly attractive feature, allowing for making predictions of, for example, text coverage based on a language corpus. A different approach would be to prioritize reliable separation of test takers on the measured construct, without necessarily being bothered about word frequency at all. Any vocabulary items that discriminate well and contribute to reliability (i.e., items with excellent psychometric properties) would be useful. To achieve optimal discriminatory power, the test would preferably be composed of words that are targeted to the individual test taker's ability level, thus necessitating a computer adaptive test (cf. Bokander, forthcoming). It is, of course, possible to combine a content related approach and psychometric considerations, as was done in Beglar and Hunt's (1999) revision of the Vocabulary Levels Test, or by the creators of the LexTALE (Lemhöfer & Broersma, 2012). In these studies, words were selected based on both frequency and careful item analysis (but not tailored to individual test takers). I am not aware of any vocabulary test that is purely psychometrically designed, and computer adaptive. Such a test would arguably be ideal for individual differences research.

4.3 Cognitive aptitudes for vocabulary acquisition

There seem to exist at least two very different aptitudes for developing an L2 vocabulary. The first is related to memorizing words intentionally, and the second is related to an incidental (statistical, implicit) process of strengthening the knowledge of each item through repeated encounters in language use. One would expect sensitivity to statistical regularities and word frequency information to be an important determinant of L2 vocabulary acquisition, at least after initial exposure that may require a more conscious learning effort (Ellis, 2002). Some evidence even suggests that words can be picked up without first having been explicitly studied (Walker et al., 2020). The validation study of the HiLAB (Linck et al., 2013) did not include vocabulary as a dependent variable, but they found significant positive correlations between an implicit serial reaction time task and tests of L2 listening and reading.

There is, as yet, limited research on effects of implicit aptitude for adult second language vocabulary development, but there is ample evidence from L1 development in children demonstrating a crucial role for statistical learning of

vocabulary (Erickson & Thiessen, 2015). It is thus reasonable to believe that tasks used in research on implicit statistical learning potentially could serve as aptitude tests to predict L2 vocabulary acquisition. A large number of tasks intended to measure various aspects of implicit or statistical learning have been developed (Siegelman, Bogaerts, Christiansen, et al., 2017) but it is not clear what kind of statistical learning tasks would predict a particular domain of SLA such as vocabulary development. Also, many tasks in statistical learning research were developed to detect group mean differences and may not have the psychometric qualities required to reliably detect individual differences in language aptitude research. In the experimental tradition, individual variability often constitutes measurement error to be avoided as far as possible. Hence, such tasks may yield low reliability and underestimate correlations when employed in aptitude research where individual differences are paramount (Cronbach, 1957; Hedge et al., 2018; Siegelman et al., 2017).

Evidence also suggests that both executive working memory (WM) and phonological short term memory (PSTM) capacity, such as non-word repetition ability, are predictive of vocabulary learning (Baddeley et al., 1998; Gathercole, 2006). Tests of WM and PSTM are considered to tap explicit cognitive processing, and could thus be expected to be more related to intentional vocabulary acquisition. For example, Martin and Ellis (2012) found moderate correlations between vocabulary scores and measures of WM and PSTM in an experiment where participants were exposed to an artificial language for an hour. Most studies documenting positive association between WM/PSTM and long term, natural language learning, have not focused specifically on vocabulary. Exceptions are, for example, Speciale, Ellis and Bywater (2004), and Service and Kohonen (1995), both reporting moderate correlations between phonological short term memory and vocabulary development over extended periods.

With traditional aptitude tests, correlations with vocabulary scores have not been impressive, but as noted above, there are only a few language aptitude studies that have included vocabulary tests as dependent variables. Existing aptitude test batteries may also not be optimally constructed to detect vocabulary development. A meta-analysis of aptitude effects in Li (2016) found that language aptitude, measured as a full test battery composite score, had a weak average correlation (r = .15) with vocabulary knowledge. The highest correlation with aptitude sub-constructs was found with phonetic coding (r = .38). Evidence in support for phonological ability as aptitude for vocabulary development was also found in a recent study (Lambelet, 2021) in which LLAMA D and E (both tapping phonological aspects of language) were significantly related to higher scores on lexical diversity measures obtained from L2 oral narrative samples.

Interestingly, tests of rote memory such as MLAT 5 or LLAMA B have tended not to produce high correlations with vocabulary measures (Li, 2016)

although these tasks are often referred to as (explicit) vocabulary learning tasks. Instead, studies examining the role of phonetic coding, phonological short term memory, and implicit learning lend support to the view that vocabulary acquisition to a large extent involves other mechanisms than just rote learning. Such findings are consistent with a connectionist view of statistical word learning, suggesting that each encounter with a word strengthens its accessibility for language use and prevents decay from memory. Measuring aptitude for vocabulary learning should thus include both tasks that tap implicit, statistical processes as well as explicit tasks requiring phonological short term memory.

A final point is that, as the review of aptitude tasks above demonstrates, few language aptitude tests are concerned with acquiring meaning in a way that resembles natural language learning. The CANAL-FT test was an exception because the whole test is built around an artificial language and the vocabulary items in the test are semantically related. However, very limited research was carried out with this test, and it is yet unknown to what extent it would be able to predict L2 vocabulary acquisition.

5 Methodology

This section discusses methodological features common to the three empirical studies included in the dissertation.

5.1 Participants, data collection and ethical considerations

The participants in study 3 (Bokander, 2020) were a subsample of those in study 2 (Bokander & Bylund, 2020). Study 4 (Bokander, 2016) used an entirely different sample. In total, 640 individuals contributed with data to the three empirical studies. They are described in the respective studies, but common to the three studies was that the participants were university students, and they were recruited via their teachers or faculty staff at the respective university and so they were largely anonymous to the researcher. By this means of recruiting, it is possible to include a large number of participants but a disadvantage is of course that the researcher has limited knowledge about who the respondents are, if they were using external help to solve test items, or if they genuinely did their best, or if they lacked motivation to perform the tasks. These are all factors that influence the score reliability, but it was hoped that larger sample size would compensate for potential inconsistencies by averaging out measurement errors that would occur due to the above mentioned factors.

The data collection was conducted in accordance with the principles of the Swedish Research Council (2017). The data used for study 4 and most of the data used in study 2 were obtained without any additional personal information about the participants, meaning that individual identification was not possible even for the researcher. Only in study 3 it was necessary to keep track of the participants between the two testing occasions but no data of sensitive personal nature was obtained. The participants reported knowledge of background languages of which none could be considered to have particular ethnic connotations, being widely spoken around the world. Informed consent was obtained from those who agreed to participate and upon the final session, they were rewarded with movie tickets as a sign of gratitude. The test sessions began with a brief presentation of the research project and the participants were informed that they were free to abort participation whenever they desired, and that the data would be coded into anonymized spreadsheet forms.

5.2 Instruments

In the papers comprising this thesis, in total seven different data elicitation instruments were employed (including the four LLAMA subtests). Although they were described in detail in the respective papers, and also to some extent in the background sections above, Table 1 summarizes them briefly here. The LLAMAs used an unfamiliar test language; the other tasks were performed in Swedish.

Table 1. Elicitation instruments used in the empirical studies.

Instrument	In study no.	Task description
LLAMA B	2, 3	Memorize and then recall the written names of pictures
LLAMA D	2, 3	Recognize previously heard spoken phrases
LLAMA E	2, 3	Associate spoken sounds with their written symbols
LLAMA F	2, 3	Learn lexico-grammatical features by studying a set of pictures and their descriptions
SweLT	4	Fill in gaps in sentences by selecting a word among four alternatives.
C-test	3	Complete truncated words in four short texts for L2-beginners.
Verbal report procedure	2	Describe (or "think aloud") thoughts and strategies during, and after, completion of LLAMA subtests.

5.3 Data analysis

The obtained data was analysed with methods that are mostly well known and widespread in quantitative behavioral research, most notably correlational methods. Information about correlation and covariance (i.e., unstandardized correlation) is used for many purposes in classical test theory, for instance in item analysis when computing item discrimination indexes (as in Bokander, 2016; Bokander & Bylund, 2020), or for internal consistency estimates like coefficient alpha. The latter is commonly conceptualized as a function of the average inter-item correlation in a set of test scores. Principal components analysis (Bokander & Bylund, 2020) starts out with a matrix of correlations or covariances, from which relational patterns in a data set are extracted.

Correlation is also used for quantifying relations between a set of scores and external criteria (i.e., other scores or ratings from tests or questionnaires). This was done in Bokander (2016; 2020), either using raw correlations or with regression analysis. The latter method considers covariance between the independent variables (in this case, the LLAMA subtests) in order to determine the unique contribution of each variable in explaining variance in the external criterion (e.g., a language test, as in Bokander, 2020).

Non-correlational methods in the thesis were the Rasch analyses (Bokander, 2016; Bokander & Bylund, 2020) and the verbal report method used in Bokander and Bylund (2020). Rasch analysis is an item-response theory (IRT) based method that takes its starting point in the probabilities of individual response patterns in the data set, from which item statistics and person ability estimates are computed (for this reason, IRT methods are often referred to probabilistic methods). Whereas the above mentioned methods are clearly associated with a quantitiative research paradigm, the thesis also contains some qualitative analyses. These are, first, the verbal report method (or 'think alound protocol') in Bokander and Bylund (2020) and, second, parts of the item content analysis applied in the same study.

6 The individual studies

This section presents summaries of the individual studies that are included in the thesis. Due to its general nature, the conceptual handbook chapter 'Psychometric assessment' (Bokander, forthcoming) comes first, thus laying the foundation for much of the methodological work in the three empirical articles that follow. This is followed by the two papers that, from their different angles, contribute validity evidence for the LLAMA aptitude tests, that is, the independent variable that is supposed to tell us something about language acquisition. The reader may then want to keep in mind the validity framework referred to in section 2.1 above (detailed in Bokander & Bylund, 2020), because Bokander (2020) takes off just where Bokander and Bylund (2020) leaves the reader, that is, at the level of extrapolation in the validity framework. Then I turn to the outcome variable in research on individual differences in SLA, that is, language acquisition. It is represented here by the vocabulary test that was developed in Bokander (2016).

6.1 Bokander (forthcoming)

Psychometric assessment. To appear in: S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge Handbook of Second Language Acquisition and Individual Differences*. Routledge.

The chapter was specified by the volume editors to comprise four sections: Overview, Technical Features, Contributions to ID Research, and Future Directions. The overview contains a brief introduction to the topic and a rationale for taking a psychometric approach in the measurement of individual differences in SLA.

The Technical Features section constitutes the main part of the chapter. It is organised as follows. First, the main steps in psychometric test development are outlined, including construct definition; item writing following specifications; pre-testing and piloting; item analysis and revision; field testing, and examination of reliability and validity. Then follows a subsection on item analysis from three theoretical perspectives – classical test theory (CTT), item response theory (IRT) and the common factor model. In particular, the discussion focuses on item discrimination which is arguably the most crucial item feature in ID research because well discriminating items are the building blocks that enable tests to reliably separate individuals on the latent trait under investigation. Without ability separation of individuals, true correlations with other variables may go undetected. The subsection closes with a brief comment on distractor analysis and differential item functioning.

Following the subsection on item analysis, some common issues in test scoring are discussed. These include how to handle situations when guessing is possible (e.g., in multiple choice tests); when items consist of stimuli that are nested in trials, as is commonly done in working memory experiments; and the relationship between observed scores and latent scores in IRT models or factor analysis.

Next, the chapter addresses reliability estimation mainly from a CTT perspective but also including a brief mentioning of corresponding concepts in IRT. The true score model is introduced, in which an observed score is interpreted as a true score plus a random error component, and reliability is defined as the ratio of error variance to observed variance. Then, reliability estimation from consecutive test administrations and from a single administration are discussed. The latter is far more common in SLA research, which usually reports the coefficient alpha which is an internal consistency estimate that is appropriate to use when the scores are unidimensional (i.e., the items target one and the same construct). Some critique against the widespread use of alpha is noted and an alternative approach, McDonald's omega, based on a common factor model (McDonald, 1999), is mentioned. This is followed by a discussion of some test score features that tend to increase or decrease reliability, including test length, between subject variance, and factors that increase measurement error (e.g., malfunctioning items, unmotivated test takers, or unclear test instructions). Some guidelines are then provided on how to report reliability in a research paper. It is pointed out that a reliability estimate is not a feature of a test, but of test scores, and care should be taken if one reuses reliability coefficients obtained in a different sample (e.g., from published test manuals). Two often cited benchmarks for reliability coefficients are provided as well as a note on how to use the reliability estimate to compute confidence intervals for individual scores. Finally, the subsection on reliability closes with a brief explanation of the IRT analogue to CTT reliability, that is, the test information function which provides reliability information along the latent ability continuum, instead of just one estimate for all test takers. Such detailed information about measurement precision, however, requires large sample sizes to produce accurate model fit, and is often not a viable option in small scale research projects.

The last part of the Technical Features section of the chapter contains an introduction to validity, a central theme in most books on test theory. The traditional, tripartite explication of validity as content related, criterion related and construct related validity, is contrasted with a more recent unitary framework of validation (put forward by, e.g., Kane, 2006). Some common methods for evaluating validity are mentioned, including content expert judgements and methods based on correlations (e.g., regression and factor analysis). The unitary framework entails a methodological extension, because it subsumes the three traditional kinds of validity together with other validity

evidence (e.g., reliability and implications for test stakeholders), and draws on a wide range of qualitative and quantitative methods for evaluating test use validity.

The third section of the chapter, Contributions to ID Research, contains a brief review of studies on working memory and language aptitude, in which a psychometric approach is highlighted with respect to some of the themes introduced in the Technical Features' section. Examples include the development of a Chinese language aptitude test where the authors reported in detail from a Rasch item analysis; the analysis of item functioning and reliability in Bokander & Bylund (2020); the use of latent factor scores in criterion validation of memory span tests; and the construct validation of the CANAL-F language aptitude battery.

The fourth and final section of the chapter points out areas in which future research on IDs in SLA could make important contributions. It is observed that few studies have reported details about the tests they employ, such as item characteristics, reliability, or construct validity evidence, and most often, the validity of the test use (in the unitary sense) seems to be simply taken for granted. To remedy this situation, researchers are encouraged to develop new measurement instruments which would allow for more latent factor studies of IDs, and also to join the current open science trend in SLA by making tests and datasets available on public repositories such as the IRIS database (Marsden et al., 2016).

6.2 Bokander and Bylund (2020)

Probing the internal validity of the LLAMA language aptitude tests. *Language Learning, 70* (1), 11–47.

Introduction

The aim of this study was to examine the LLAMA language aptitude tests with respect to internal aspects of validity, that is, item functioning and reliability, and also to examine some evidence of construct validity (relations between subtests and test takers' response behavior). The background was as follows. During the recent decade a growing number of studies have been published involving the construct of language aptitude, drawing on data from the LLAMA language aptitude tests (Meara, 2005). The different areas of research include age related effects in SLA (e.g., sensitive periods for L2 learning and L1-attrition); the role of aptitude at different L2 proficiency levels, aptitudes for explicit and implicit learning, instructed L2 learning and feedback, naturalistic SLA abroad, aptitude in relation to other cognitive constructs (e.g., working memory, intelligence, musical aptitude), oral language proficiency, and neuro-cognitive studies. Findings from these studies have the potential to impact our accumulated understanding of SLA processes. However, as pointed out by the

creator of the LLAMA, these aptitude tests had not undergone any substantial validation before being published and the LLAMAs should not be used in high-stakes situations (Meara, 2005).

Addressing the lack of validity evidence for the LLAMA tests, we took as our point of departure a validation framework proposed in the educational measurement literature (Kane, 2006) and applied to second language research in Purpura, Brown and Schoonen (2015), although the present study did an adaptation specifically to cater for the needs of language aptitude tests. The validation framework (laid out in detail in the paper) is built around a chain of inferential links, or levels, going from test internal validity evidence at the single item level, via subtest-level to the whole test battery level and its relationship to behavior that the aptitude test sets out to predict. Investigation of item characteristics and internal consistency requires access to all test responses and not just total scores. Because the LLAMA does not provide scores at item level (only subtest total scores are recorded), we developed a replica of the test that could be administered in a web browser and that gave us access to each individual item response from each test taker. Only a few previous studies have reported reliability coefficients for LLAMA scores and most of these coefficients have been in the lower range for what is usually deemed acceptable reliability, meaning that the scores are likely to contain a lot of measurement error.

Three research questions guided the study: first, to what extent the individual item scores are reliable [RQ1], second, to what extent the subtest total scores are reliable [RQ2], and third, to what extent the entire LLAMA test battery may reliably reflect a latent aptitude construct [RQ3]. The first of these questions relates to the scoring inference of the interpretive argument outlined in Purpura, Brown & Schoonen (2015); the second relates to the generalization inference and the third to the explanation inference in the validity framework. The interpretive argument also includes inferences of extrapolation and interpretation, which was not evaluated in the present study because such an investigation would necessitate data on language learning outcomes as well.

Method

The LLAMA consists of four subtests, three of them containing 20 items and one (LLAMA D) containing 30 items. To handle measurement error, a large dataset was required to address the research questions. Data collected on different locations were aggregated and in total, complete score sets from 350 informants were used in the study. The first question was answered with classical test theory (CTT) and item response theory (IRT), by computing proportion correct responses, discrimination and Rasch fit indexes for each item in the LLAMA. The second question was answered with CTT reliability estimates (internal consistency) and overall Rasch model fit, and the third question was explored by means of principal component analysis (PCA),

content analysis, and response time analysis. PCA is a method for dimension reduction by identifying patterns in the correlations between variables, similar to exploratory factor analysis.

Results

The analysis for the first research question revealed that many items produced less than satisfactory discrimination properties and Rasch item fit, and this was particularly so in subtest LLAMA D, in which almost one third of the items performed near random. LLAMA F also produced item statistics that were less than ideal, and one item was found to be wrongly coded, thus awarding zero points to test takers who actually got it right. Only subtest LLAMA B produced reliable item statistics both under the CTT and the Rasch paradigms. At the subtest level (the second research question) we found that reliability and Rasch model fit was lower in the subtests which contained more non-discriminative items, in particular LLAMA D. The finding is generally in accordance with other studies in which LLAMA D consistently has produced low reliability estimates (e.g., Gisela Granena, 2013a). Finally, the findings related to the third research question corroborated the whole-test structure found in Granena (2013) in which subtest LLAMA D loaded on a separate principal component than the other three subtests. Support for a three-dimensional aptitude construct, such as Skehan's suggestion that aptitude consists of phonological ability, memory and analytic ability, could not be found in our dataset, meaning that LLAMA may not be sufficiently effective in distinguishing between different aptitude dimensions. The content analysis included PCA of items in the subtests D and F (those with internal consistency), in order to figure out possible reasons for this lack of consistency. In LLAMA D, items seemed to cluster according to their familiarity status (i.e., new or familiar stimuli in the initial practice phase of the test). In LLAMA F, items seemed to cluster according to grammatical content, thus contributing to lower internal consistency. The subtest LLAMA E has been relatively easy in many studies, including ours, with near ceiling effects (a phenomena that may reduce correlations with other variables). Our analysis of content and response times suggested that this was because some items in the test allowed for strategies for solving the items without engaging the actual skill that the test sets out to measure.

Implications

In short, the results suggest that there is potential for improving the LLAMA test battery. Only subtest LLAMA B fitted well to the item response (Rasch) model and displayed no odd item behavior, which is most likely a consequence of the test format – unlike LLAMA D, E and F, subtest B does not use a binary response format, thus mitigating the impact of correct guesses introducing noise in the data. Our recommendation to researchers is to use the LLAMAs with

proper care when interpreting test scores and, as pointed out by Meara (2005) not using LLAMA in high stakes situations.

6.3 Bokander (2020)

Language aptitude and crosslinguistic influence in initial L2 learning. *Journal of the European Second Language Association, 4*(1), 35–44.

Introduction
The study was intended to examine predictive validity evidence for the LLAMA language aptitude tests among beginner learners of L2 Swedish, also taking into consideration the learners' L1 and possibility to transfer. Previous research has yielded mixed findings with this aptitude battery regarding correlations with L2 outcomes and some of them are reported in Bokander & Bylund (2020). Most relevant as a background for this study was research demonstrating a relationship between sound sequence recognition (LLAMA D) and overall language gains among beginners (Artieda & Munoz, 2016). That finding was particularly interesting because it aligned with Skehan's (1998; 2019) prediction that phonological processing (supposedly targeted by LLAMA D) is crucial to the earliest stages of learning a language. The present study thus entertained the hypothesis that the aptitude trait measured by LLAMA D would be implicated in the overall gains by the participants [RQ3]. Because the participants came from mixed L1 backgrounds, it was also believed [RQ2] that high aptitude would be most beneficial to learners with typologically distant L1s, similar to how aptitude has been shown to be more beneficial for late age-of-onset learners than for early bilingual learners (Abrahamsson & Hyltenstam, 2008). The rationale behind this idea was that aptitude may serve as a catalyst in particular for learners that are facing a greater learning challenge (older starters, or speakers of typologically distant L1s). Finally, the literature on individual differences in language learning has often compared the relative impact of various ID constructs on SLA, such as aptitude, motivation, personality, learning styles, etc. This study sought to add to this list of relative strength issues by comparing the advantages of having high aptitude, or having a typologically similar L1. This was explored as [RQ1] in the study.

Method
Ninety-two international students learning Swedish at a Swedish university took part in the study. They were studying various subjects at the university and took part in the Swedish course out of interest in learning the local language and culture and not as a part of their main study program (however, they did receive course credits upon successful completion of the course). The students had different language backgrounds, and about half of them had a Germanic L1

(typically, German or English). Among non-Germanic L1 speakers, Mandarin and Japanese were the most represented languages.

The Swedish course is traditionally offered every semester to new students and it was a well established observation among the teachers that European students, in particular from Germanic speaking countries, typically outperformed students with typologically more distant L1s. In the present study about half of the participants were speakers of a Germanic L1. At an early point during the 5-week long introductory Swedish language course, those who desired to participate in the study completed the LLAMA language aptitude test battery. At the end of the course, they completed a C-test constructed for the purpose of the study, based on easy texts from various textbooks for beginners. The C-test (Klein-Braley, 1997) technique is based on the idea of reduced redundancy (similar to the cloze procedure), with the second half of every other word deleted. Morpho-syntactically, the language in the C-test was very basic so the greatest challenge for the test takers may be assumed to be related to vocabulary.

Results

The scores from the C-test were regressed on the LLAMA scores in a multiple regression analysis and standardized beta-coefficients were computed. In the full sample (N = 92), of the four aptitude subtests only LLAMA D displayed a significant but small effect. The effect of L1-group was large and significant, indicating that language background clearly had more predictive power than aptitude scores. Upon inspection of the C-test scores and raw correlations, it was obvious that the difference in Swedish achievement in the Germanic speaking and the non-Germanic speaking groups performed very differently on the test – the former clearly outperforming the latter. Hence, separate regression analyses were carried out in the two L1 categorized subsamples (typologically close versus distant). The results showed that aptitude scores significantly predicted L2 outcomes in the typologically close (Germanic L1) subsample. Out of the four LLAMA subtests, it was those involving sound processing (LLAMA D, and to a lesser extent, LLAMA E) that were significantly related to L2 outcomes. The finding is consistent with the theory of differential aptitude effects at developmental stages proposed by Skehan (1998) and also with previous research (Artieda & Muñoz, 2016). However, the fact that no effect was found in the typologically distant subsample suggests that learners may need to progress to some point above complete beginner level for language aptitude to take effect. Importantly, the C-test score variance was about the same in both subsamples, indicating that the lack of aptitude effect in the non-Germanic L1 group was not due to a statistical floor effect mitigating any correlations due to low variance (which may sometimes be the cause of a null finding in correlational analysis).

Implications
A main implication of the study is that crosslinguistic influence may be of greater importance than language aptitude in mixed-L1 student groups, at least in the initial phase of L2 acquisition. The C-test used as L2 criterion in this study could be correctly completed with very basic grammatical knowledge, but the lexical load was probably high for this learner level. Test takers with a related L1 may draw on similarities such as cognates, which would provide a large benefit compared to test takers whose language background does not permit positive transfer. However, it was also observed that with a 'levelled playing field' (i.e., when test takers with similar language background were compared), and for test takers that were in a position to draw on positive transfer, it was clear that phonological aptitude was related to initial L2 achievement – at least when the challenge for the learners mainly consisted of remembering vocabulary to complete the gaps in the test.

6.4 Bokander (2016)

SweLT 1.0 – konstruktion och pilottest av ett nytt svenskt frekvensbaserat ordförrådstest. *Nordand*, 11(1), 9–30. [SweLT 1.0 – construction and piloting of a new, Swedish, frequency-based vocabulary test]

Introduction
Noting the lack of a widespread and reliable Swedish vocabulary test that could be used for research and placement purposes, the aim of this study was to pilot a discrete point, multiple choice, receptive vocabulary test based on word frequency. Inspiration for the Swedish vocabulary levels test, SweLT (Bokander, 2016) came mainly from the Vocabulary Levels Test, VLT (Nation, 1983; Schmitt et al., 2001) and the Vocabulary Size Test (VST, Nation & Beglar, 2007). These tests are based on frequency ranked word lists, derived from large language corpora assumed to be representative of the language for which the test is intended. Notably, there exists a relatively old corpus linguistic tradition in Sweden with early frequency lists being published around 1970 but this work does not seem to have made its way into language testing or education.

Thus there appears to exist a lacunae within Swedish language testing that SweLT was designed to fill. Four research questions were posed. The first and second questions concerned the internal validity of the test [RQ1], that is, item properties (difficulty, discrimination) and the relationship between word frequency and difficutly [RQ2]. The third question was related to external validity and investigated the association between SweLT scores and self or teacher rated Swedish proficiency level. Finally, it was investigated to what extent SweLT scores may reliably estimate learners' vocabulary size [RQ4].

Method

Because the accessible frequency based word list that I decided to use (Forsbom, 2006) was limited to about 8,500 lemmas, it was not possible to include a 10K level (as in the VLT), or for that matter, to aim for a Swedish replica of the VST with its 14 K-bands. The word list (Forsbom, 2006) was sampled at the 2K, 3K, 5K, and 8K frequency bands with 20 words from each band, resulting in 80 target words. Only content words were included (nouns, verbs and adjectives) because the rationale behind the development of the SweLT was primarily to target semantic aspects of vocabulary and not syntactic aspects (although grammatical words certainly play an important role in vocabulary development as well). The words were sampled approximately in the proportions that they were represented in the word list (an approximate ratio of 5:3:2 between nouns, verbs and adjectives).

In the choice of item format, although inspiration for the SweLT came from the VLT and VST, it was decided to abandon the matching formats used in those tests (the VLT and the VST actually use quite different item formats, but both rely on the idea of matching written meanings with target words). One reason for this was the difficulty in creating a definition for each distractor. Instead, a multiple choice completion format was used, that had been reported as well functioning in a study of TOEFL vocabulary items (Henning, 1991). Items were created in the following way. A target word was embedded in a short sentence of high frequency vocabulary, with special attention to creating plausible collocations with the target word. To this end, a Swedish corpus tool (Borin et al., 2012) was consulted and the most common collocational contexts were used as inspiration during the creation of the sentences carrying the target word. This resulted in items in which the target word was used in its most 'normal' way, thus avoiding to confuse test takers with uncommon word usage. Finally, after creating a well formed context, the target word was excluded from the sentence and placed together with three distractors. The latter were chosen so that they would be grammatically correct alternatives, but produce a nonsensical sentence. The distractors were also chosen from lower frequency bands (i.e., more difficult) than the target word to minimize the probability of the distractor being known to the test taker and thus easy to eliminate (i.e., improving guessing odds). The test form was administered online via the internet or in paper-and-pencil format by the participants' teachers.

Results

The 3K and 5K frequency bands worked very well for the sample of participants, in the sense that item difficulties and discrimination were clearly satisfactory or very good. The 2K frequency band was too easy to yield meaningful measurement in the sample and would thus need to be tried out with a different sample of lower ability. However, the Rasch analysis indicated that the items were not necessarily flawed, in contrast to the sample dependent discrimination index which indicates weak discrimination when test taker

ability and item difficulty are poorly matched. The 8K level showed some signs of needing further revision, with several items having unsatisfactory parameter values both under the classical and the Rasch measurement models. The reliability (coefficient alpha) was good in all frequency bands except the 2K level, which was unsurprising given that alpha tends to be lower when there is little variance and low discriminatory power in the scores (low inter-item correlations). On the whole, the internal validity of the SweLT scores in the study thus seems good, in particular in the 3K and 5K bands. The external validity was examined by correlations with self or teacher reported proficiency and as expected, a moderate to strong relationship was detected at group level. However, because of the large spread of scores at each proficiency level, any predictions for individual learners would have large errors. Similarly, when investigating whether the SweLT could provide some kind of vocabulary size measure, the findings were in line with other research in this area but the error margin when extrapolating from a SweLT score to an individual's receptive vocabulary size would be substantial.

Implications
The SweLT thus seems to be a reliable indicator of receptive vocabulary in particular for learners at intermediate level, but further studies are needed to establish its relationship with external criteria such as the CEFR scale. For estimation of absolute vocabulary size, the SweLT does not seem to have enough precision. Some items, mainly at the 8K frequency level, will also need revision in future versions of the test.

6.5 Summary of the results

The results from the three empirical studies that provide validity evidence for LLAMA and SweLT may thus be summarized as follows. Study 4 examined a pilot version of SweLT, and as expected, some items would need to be revised before this test can be employed in language aptitude research. The high frequency level (2K) was inadequate for separating test takers' vocabulary knowledge. Levels 3K and 5K seem to work very well for this purpose. SweLT also seems to possess external validity, as it significantly separated learners at different communicative proficiency levels, although as one may expect, there was substantial overlap between groups. The validity of LLAMA was explored in study 2 (internal validity) and study 3 (external validity). Only LLAMA B displayed item properties and overall reliability that would be acceptable for high stakes testing. The other subtests were to a greater or lesser extent associated with low reliability at item and test level, as well as dubious construct representations. LLAMA D showed two distinct dimensions, which poses a problem for score interpretations from this test. LLAMA E seems to tap more analytical skills than previously supposed, and several items in LLAMA F were

inconsistent with the rest of the scale, thus adding measurement error. However, study 3 showed that LLAMA behaved as expected with respect to the role of auditory processing among beginners, and this finding was also in line with earlier research. LLAMA may thus be considered useful, given that findings from this test battery are interpreted with some care.

7 General discussion

This thesis aimed at exploring to what extent LLAMA and SweLT may serve as valid instruments in research on aptitude for vocabulary acquisition, based on the premise that vocabulary size is a convenient proxy for general L2 proficiency. As the review of the literature shows, vocabulary has been remarkably absent in language aptitude research which is surprising given the major role that vocabulary has in language comprehension and use. One possible explanation for the modest interest in vocabulary among aptitude researchers could be that in the times when the MLAT (Carroll & Sapon, 1959) was developed and during the decades that followed, language learning research was largely preoccupied with grammar while vocabulary was a neglected topic (Meara, 1980). Another possible explanation is that the few aptitude studies that have included vocabulary as an achievement criterion, have not found very impressive correlations (Li, 2016). Such findings may, as noted in Section 4 above, not necessarily be due to a lack of relationship between language aptitude and vocabulary. It could equally well have a methodological cause, if aptitude tests are not designed to tap vocabulary acquisition and/or vocabulary tests are not sensitive to individual differences (i.e., unreliable). This section begins by discussing validity evidence in support (or not in support) of a research design using LLAMA and SweLT to investigate aptitude for vocabulary learning. This is followed by a few more general remarks with relevance for the present thesis and future test use.

The evidence related to criterion validity (study 3) suggests that LLAMA D may have some promise for detecting individual differences in vocabulary acquisition. One earlier study found positive correlations between LLAMA D and a set of lexical measures (Granena & Long, 2013), albeit with only marginal significance due to the small sample. Assuming that the C-test in study 3 (Bokander, 2020) was tapping vocabulary skills, the result lends support to the finding by Granena & Long. It thus seems as if LLAMA D (yielding the highest correlation in both these studies) or a similar task with improved psychometric properties, would be a potential candidate to include in aptitude research directed towards vocabulary development. LLAMA B, which is referred to as a vocabulary learning task, did not predict any outcome variance in study 3 which might seem odd if one assumes that the C-test was mainly a test of vocabulary knowledge. However, this finding may be due to the two aptitude subtests (LLAMA B and D) targeting very different aspects of vocabulary acquisition. LLAMA B is a rote learning task similar to learning words incidentally from a word list or flashcards. LLAMA D, on the other hand, might tap into more implicit processing involved in building a vocabulary over some time. As noted in the literature review earlier, previous studies have suggested a link between LLAMA D and implicit learning or implicit memory, although the findings so

far are inconsistent (Granena, 2019). More research is clearly needed to figure out exactly what the LLAMA D measures and its relationship to vocabulary acquisition.

Turning to the internal validity of LLAMA, Bokander and Bylund (2020) clearly confirmed the low reliability issue with LLAMA scores found in several other studies and this particularly pertains to LLAMA D, which has not produced internal consistency coefficients above .65 in any study to date. However, two observations from LLAMA D scores merit some further thought. First, in spite of the low reliability coefficients (alpha) found with this test, it has still produced moderate correlations with L2 learning criteria. Second, our factor analysis in study 3 found that LLAMA D is probably not a unidimensional test, because items of the 'familiar' type and the 'new' type were loaded on separate factors. In particular the 'new' items produced noisy data, suggesting that it is very difficult for participants to accurately report 'new' items as never encountered before. Thus, lack of unidimensionality could potentially make coefficient alpha a poor estimator of reliability for LLAMA D, because alpha assumes a common test factor with equally discriminating items (cf. Bokander, forthcoming). If an instrument is composed of one part that is unreliable or random, and another part that may be highly reliable, an overall reliability estimate may turn out lower due to the random part (noise). However, it may still be able to produce correlations (albeit weaker) with an external criterion. Some support for this line of reasoning comes from the test-retest correlation (a different method for estimating reliability) reported in Granena (2013) which was not worse for LLAMA D than any of the other LLAMA subtests. More research is clearly needed but it seems unmotivated to exclude LLAMA D in research (as suggested in Li & Zhao, 2021) just because of the low internal consistency estimates it tends to produce.

It was observed in the literature review above that the MLAT has yielded among the highest correlations with language learning outcomes in aptitude research. Two features of the MLAT that are different from LLAMA may have contributed to the superior predictive performance of the MLAT: a more reliable test format (longer tests, more response options) and the inclusion of L1 related content. The first of these features is methodological and should be considered if an adaptation of a LLAMA test be used, as in Suzuki & DeKeyser (2017), or a new similar test be developed. To some extent this issue has recently been addressed in the development of the new LLAMA tests currently being tried out as beta versions (Meara & Rogers, 2019). No research has yet been published on these tests so it is unknown to what extent they will be able to address the methodological shortcomings in the original LLAMAs. The second feature is content related and has to do with the extent to which an aptitude test should draw on L1 ability. It is often pointed out as a particular advantage of the LLAMA that it is 'language neutral' and Rogers et al. (2017) indeed found that the LLAMA tests seem to work equally well with participants

of different L1s as long as they are familiar with the latin alphabet. Paradoxically, this feature could have contributed to the lower predictive performance of the LLAMA in comparison with the MLAT, because, as seen in the literature review above, strong evidence suggests that L2 aptitude is linked to L1 ability. As seen in the review of aptitude item types above (section 3.2), in particular MLAT 3 (Spelling Clues) seems to tap into the participants' L1 skills (identify misspelled words and finding their synonyms in English). Hence, it is possible that an aptitude test that is completely void of all L1 influence would not do its job very well. A possible direction for future aptitude research would be to adapt a more contrastive approach, and design aptitude tests tailored to the L1 of the participants, rather than seeking to be language neutral.

Turning to the issue of how to quantify L2 knowledge , the Swedish levels test, SweLT, was designed to provide researchers or educators with a rough estimate of test takers' receptive vocabulary knowledge. Study 4 described the piloting of this test and the first three research questions were directly related to the validation framework introduced in the introduction to this thesis. They concerned the stages in Kane's (2006) model labeled generalization (item functioning and reliability), explanation (word frequency as a predictor of item difficulty) and extrapolation (the relationship between SweLT scores and the CEFR levels). Taken together, the results from study 4 (Bokander, 2016) suggest that a refined version of SweLT, after revision of some items that did not perform as expected, could have the potential of being a useful research instrument. Reliability was found to be acceptable and extrapolation to a criterion variable (CEFR level) was possible though not highly precise. A couple of issues would need to be addressed, however, before using this test in an individual difference study of language aptitude effects. The first is that it is yet unknown how well SweLT would work with lower level learners. There was a distinct floor effect in the 2K frequency band, rendering this level useless for detecting individual differences in vocabulary knowledge among the (mostly intermediate level) learners that took part in the pilot study. This level thus needs to be further piloted with less experienced learners. It was, however, judged to be too difficult to include in study 3 (Bokander, 2020) with absolute beginners, in which case an easy C-test with more predictable psychometric properties was believed to yield more reliable information about L2 proficiency. A second point when discussing the possible role of SweLT in an individual differences study is that the aim of testing is rather different from the educational aims behind frequency band based vocabulary testing. Using tests based on frequency bands may inform educators about, for example, what kind of reading would be most appropriate for the students, or the kind of texts they would be likely to find too difficult because their vocabulary does not provide enough text coverage. Research on individual differences, on the other hand, is concerned about maximizing variance among the participants. It may be that

this is not done best with a frequency based approach. Rather, with a purely psychometric approach aiming at maximizing item discrimination and reliability, but sampling test words with some other method than from frequency lists (for example, random sampling from a dictionary), could be a better, or at least an alternative, way to proceed. The LexTALE vocabulary test, developed for psychological research (Lemhöfer & Broersma, 2012), was designed using word frequency as a rough guide to control the level of difficulty, after which items with the best discriminating power were selected, thus following best practice in psychometric test development. This seems like a promising approach for future refinement of the SweLT.

It would probably merit to accord some attention to the last inferential level in Kane's (2006) validation framework which concerns implications of test use in practice. Although that is not a central topic of this thesis, it would probably be a serious omission not to say anything at all about it. Therefore, I would like to end this discussion with a few words on aptitude testing outside the research context, and the hypothetical scenrario under which aptitude testing could be part of the Swedish L2 training offered to adult immigrants by municipalities and private educational organizers in Sweden. In the Introduction it was observed that language aptitude tests have been advocated as a means of tailoring language education to the individual needs of learners. It was also observed that there has been some criticism voiced regarding the efficiency of the language programs offered to adult immigrants in Sweden (Svenska för invandrare, SFI). One common point of criticism is that language courses are not individually adapted, and that there is a large variability in the rate of progress also among learners that have been assigned to a group based on their educational background (Skolinspektionen, 2018).

On the surface, language aptitude tests for placement decisions may seem like the perfect solution to this problem. Language learners would then be assigned to groups that share a similar level of aptitude and receive instruction much better tailored to their needs. There are, however, at least three great challenges for such a solution. First, using language aptitude tests for practical placement decisions in education would require highly reliable tests in order to make the decisions justified. In its present state, the LLAMA seems unlikely to be able to meet such requirements, as demonstrated in Bokander & Bylund (2020). Second, the language courses at SFI take place at beginning, up to intermediate level, with participants who pass the final exam performing at approximately level B1 of the Common European Framework of Reference (SOU, 2013:76). In Bokander (2020) it was suggested that at beginners' level, language aptitude may be a much less reliable predictor of L2 acquisition than L1 typological proximity. This suggests that it would make more sense to place beginning students according to their L1 than to their language aptitude. Third, the studies in this thesis, as well as most studies on language aptitude and SLA in general, were carried out with samples of relatively high educational level.

This is a well known limitation in much research done in SLA and other related disciplines (Andringa & Godfroid, 2019). The educational backgrounds and experiences of participants in SFI is known to vary greatly; the same classroom may include experienced professionals from a nearby EU country, alongside war refugees from a quite dissimilar culture and with limited exposure to higher education. Following suggestions by, for example, Young-Scholten (2013), research on individual differences would need to be carried out with more representative samples than has hitherto been the case, in order to figure out whether our knowledge about language aptitude generalizes over different kinds of L2 learners. There are thus many questions that remain to be answered before we are in a position to advocate the implementation of aptitude tests for placement decisions in Swedish adult L2 education.

In the introduction to this thesis, I formulated an overarching aim for this thesis and its included papers, which was to outline some key theoretical and methodological aspects of measurement practices in the study of language learning aptitude. If successfully met, the thesis or parts of it could make an important contribution to research in the field of individual differences in SLA. Theoretically, the main contribution of the thesis is arguably to emphasize a greater focus on validity issues in aptitude research. By outlining a validation framework solidly grounded in contemporary best practice in educational measurement (Kane, 2006), one would hope that more solid findings will emerge, allowing for justified decisions in particular if/when research findings are put to work in non-academic contexts. Methodologically, this thesis has only scratched the surface of all the possibilities that exist to investigate language aptitude, but one thing that I hope will be a take home message from my work is the importance of according detailed attention to the psychometric qualities of tests that are used in language research. Awareness of how individual test items work in a measurement instrument and how they contribute to sum scores and trait interpretations, should be an important part of any research endeavour in SLA, in particular when findings are of interest to policy makers, for example, in the field of education. Perhaps this contribution is particularly essential in a Swedish research context because of the more limited number of studies conducted on L2 Swedish as compared to English L2 environments. The lack of an efficient vocabulary test in Swedish was partially addressed in Bokander (2016) but more work is clearly needed in order to equip Swedish L2 researchers with high quality measurement instruments.

8 Conclusions and future directions

This thesis investigated methodological issues in using the LLAMA as an independent variable in aptitude research, thus essentially defining language aptitude operationally as whatever the LLAMA tests measure. Some studies that based their findings on LLAMA scores have several hundred citations in Google Scholar (e.g., Abrahamsson & Bylund, 2008; Granena & Long, 2013). The proliferation of these findings is thus considerable and the findings may have a serious influence on how knowledge is construed in the field of SLA. It is then worrying to find, as done in study 2, that from a psychometric point of view, the LLAMA tests leave much to be desired. To improve the situation in future knowledge building, the aptitude research community would do well in adhering to calls for increased methodological rigor, including more attention to the validity of test instruments.

The thesis also discussed the underexplored option in language aptitude research of representing L2 ability with vocabulary measures, and SweLT (after being further refined) was proposed as an alternative when the target language is Swedish. It was suggested, however, to prioritize psychometric discrimination rather than valid content sampling, in order to maximize variability among participants. If one would want to use vocabulary as a proxy for L2 ability, study designs would need to include language aptitude tasks that have the potential to tap vocabulary development. At present, no ready-to-go aptitude battery exists for this purpose. Such a set of independent variable tasks would necessarily need to include a range of measures including working memory, phonological short term memory, implicit learning, as well as tasks from existing language aptitude test batteries. As demonstrated in this thesis, LLAMA D seems to be an interesting candidate, but more research is needed to figure out the theoretical rationale behind whatever that test measures and its relationship to L2 vocabulary acquisition. In addition, it was observed that many tasks on implicit statistical learning have been developed within the experimental paradigm in psychology, aimed at reducing individual variation to a minimum, thus making them highly unsuitable for correlational studies of individual differences. When employing test instruments to function in a different research context than they were intended for, new validation and possibly a thorough revision of the tests will be needed.

A final point concerns the problem of mainly including high-educated participant samples in language aptitude research, briefly addressed above. My own studies in this thesis are no exceptions to the, seemingly common, habit in SLA to taking the easy way out and do research with convenience samples, perhaps students in the researcher's vecinity. In today's Sweden, increasingly large groups of language learners are presumably not of the kind that readily lines up to enthusiastically perform sets of cognitive tasks and language tests

for a symbolic reward in return. Studying individual differences in language acquisition among people whose financial situation, educational level, language skills and cultural habits make them more difficult to approach and involve in research that relies on obtaining reliable test scores, is an intricate challenge but one well worth pursuing in future research.

9 Sammanfattning på svenska (summary in Swedish)

Inledning

Föreliggande avhandling behandlar metodologiska frågor i forskning om språkbegåvning med fokus på de testverktyg som används för att mäta språkbegåvning (den förklarande variabeln) och uppnådd språkbehärskning (utfallsvariabeln). Språkbegåvning antas vara en av flera bakomliggande faktorer som kan förklara variation i hur snabbt och hur väl människor lär sig ett nytt språk i vuxen ålder (Dörnyei & Ryan, 2015). Modern forskning om språkbegåvning har bedrivits sedan 1950-talet och test som konstruerats för att mäta språkbegåvning har visat sig kunna förutsäga inlärares framgångar i språkinlärning med relativt hög träffsäkerhet i jämförelse med andra bakomliggande faktorer.

Det finns flera anledningar till att närmare vilja undersöka hur språkbegåvningstest fungerar och hur pålitliga data de genererar. Dylika test har åtminstone i teorin en funktion att fylla i situationer då individer placeras in i grupper inför en språkkurs. Forskare har framhållit nyttan med att ta hänsyn till språkbegåvning för att varje inlärare ska få en individuellt anpassad studiegång, vilket både skulle kunna förbättra studieresultat och göra inlärningen mer givande för individen (Robinson, 2001). I utbildningen i svenska för invandrare (SFI) används idag inlärares tidigare studiebakgrund som enda urvalsverktyg för att placera individer i olika studievägar. Problem i SFI-utbildningen har lyfts fram i rapporter och massmedia, och en återkommande punkt tycks vara bristen på individanpassning och stor variation i inlärningshastighet mellan individer på samma nivå och i samma grupp (Skolinspektionen, 2018). Vid en första anblick tycks därför språkbegåvningstest ha viss potential att bidra till bättre gruppsammansättningar i vuxenutbildningen i svenska som andraspråk. Praktisk tillämpning förutsätter emellertid att det finns pålitliga språkbegåvningstest som verkligen mäter vad de utger sig för att mäta och gör detta på ett träffsäkert sätt. Kraven på hög testkvalitet bör gälla även i forskningsstudier om språkbegåvning, där det är av stor vikt att de test som används är av god kvalitet och ger rättvisande resultat. Bristfälliga testverktyg riskerar att underminera kunskapskonstruktion inom språkvetenskap och i förlängningen få konsekvenser utanför forskarvärlden, eftersom forskningsrön om språkinlärning kan plockas upp av beslutsfattare, exempelvis inom utbildningsväsendet, och omsättas i praxis med konsekvenser för individer.

Frågan om huruvida ett test fungerar väl för sitt tilltänkta syfte brukar undersökas i validitetsstudier. Ett centralt syfte med denna avhandling är att undersöka validitet i språkbegåvningstestet LLAMA, men om LLAMA eller

liknande test på ett meningsfullt sätt ska kunna användas i svensk forskning om vuxnas inlärning av svenska behövs även tillgängliga, praktiska och valida test av svensk L2-färdighet. Av denna anledning inkluderas i avhandlingen en valideringsstudie av ett nytt test av svenskt inlärarordförråd, för att representera utfallsvariabeln i forskning om språkbegåvning.

Validering, terminologi och testteoretiska överväganden.

Validitet handlar om hur bra information ett test ger om det som testet avser att mäta och proceduren att utvärdera validitet kallas för validering. Inom psykologi och utbildningsvetenskap har begreppet validitet haft olika innebörder under de senaste hundra åren. I denna avhandling används en enhetlig valideringsmodell som baserar sig främst på Kane (2006), senare utvecklad för andraspråksinlärning i Purpura, Brown och Schoonen (2015). En enhetlig syn på validitet innebär att i en och samma valideringsmodell inordna en rad olika aspekter av testkonstruktion och testfunktion, inklusive information om enskilda testfrågor och om reliabilitet. Validering enligt Kane (2006) sker i olika nivå där varje lägre nivå är en förutsättning för validitet i en högre nivå. De logiska länkarna från en nivå till en annan kallas för inferenser och validering enligt denna modell handlar om att stärka dessa inferenser, enligt analogin att ingen kedja är starkare än dess svagaste länk. Den första nivån avser huruvida en observerad testpoäng ger en bra representation av testtagarens kunskap i förhållande till testfrågorna. Det är en inferens från enskilda responser till en testpoäng och kallas därför poänginferens (scoring inference). Nästa inferens gäller om testpoängen kan generaliseras till att gälla alla tänkbara varianter av samma test (vilket bygger på föreställningen att ett enskilt test utgör ett urval av frågor från en större mängd hypotetiskt tänkbara frågor). Denna generaliseringsinferens (generalization inference) fungerar ungefär som reliabilitetsanalys i klassisk testteori. Den tredje inferensen innebär en extrapolering från det generaliserade resultatet (som nu alltså även innehåller information om poängens tillförlitlighet) till något externt kriterium, t.ex. beteende utanför testsituationen. Detta steg i valideringen kallas extrapoleringsinferens (extrapolation inference) och motsvarar ungefär kriterievaliditet i klassisk teori (jfr ovan). Den sista inferensen som Kane diskuterar gäller implikationer av testanvändning, alltså om de totala konsekvenserna för alla inblandade (stakeholders) är övervägande goda eller dåliga. I denna avhandling dominerar undersökningen av de första inferenserna i Kanes modell (poänginferens och generaliserngsinferens), vilket medför fokus på enskilda frågors funktion och reliabilitet (s.k. intern validitet). I följande stycken avhandlas dessa båda egenskaper något mer i detalj, med tyngdpunkt på några vanliga överväganden som testforskaren måste göra.

Itemanalys

Itemanalys spelar en viktig roll i två av de empiriska studierna i denna avhandling (Bokander, 2016; Bokander & Bylund, 2020). Itemanalys görs för att säkerställa att alla testfrågor (item) bidrar på ett meningsfullt sätt till att inhämta information om testtagarens kunskap. Två metoder för itemanalys förekom i denna avhandling, klassisk analys och Raschanalys. I klassisk testteori är två mått av särskild betydelse, en frågas svårighet och dess förmåga att diskriminera mellan deltagare som har olika nivå av det som testet vill mäta. En frågas svårighet brukar anges som andelen testtagare som klarar frågan. Diskriminering brukar anges som frågans korrelation med totalpoäng på testet eller något annat relevant kriterium. I Raschanalys beskrivs testfrågor och testtagare med avseende på hur nära deras egenskaper passar den s.k. Raschmodellen (Rasch, 1960) som försöker förutsäga hur frågors svårighet och testtagares färdighet samverkar. Testfrågor som avviker mycket från modellens förutsägelse får starkt avvikande värden för modellpassning och bör därför undersökas närmare eller plockas bort från testet. Dessa "fit statistics" ger ofta en mycket bra möjlighet att diagnosticera testfrågors funktion och användes i denna avhandling som ett komplement till klassisk testanalys. I klassisk testteori är itemvärden beroende av det aktuella urval deltagare som besvarat frågorna. En i grunden väldesignad fråga kan få dåliga värden om deltagargruppen är för stark eller för svag relativt frågan. Raschanalys undviker i hög grad detta problem, vilket bland annat demonstrerades i Bokander (2016).

Reliabilitet

Inferensen som rör generalisering (Kane, 2006) handlar i hög grad om att undersöka reliabilitet. I avhandlingens empiriska studier gjordes detta enligt klassisk testteori, vilken bland annat gör gällande att reliabilitet är en egenskap hos testsvar och inte en egenskap hos testet självt. Reliabilitet i testsvar är nödvändigt både för att kunna dra valida slutsatser baserat på testpoäng och för att kunna använda testpoäng i korrelationsstudier med andra variabler. För att beräkna reliabilitet användes i denna avhandling koefficient alfa, som ger en indikation om den interna konsistensen i datasetet. Intern konsistens innebär något förenklat att alla frågor arbetar i samma riktning och bidrar med information om konstruktet som testas. Det närmast slentrianmässiga bruket av koefficient alfa för skattning av reliabilitet har under senare år fått utstå kritik inom psykologisk forskning och diskussionen har även nått andraspråksforskningen (Plonsky, 2013). Kritiken grundar sig främst på att användning av alfa för att skatta intern konsistens, vilken i sin tur tolkas som en indikation på reliabilitet, bygger på strikta förutsättningar om hur testdata är distribuerade (Dunn, et al., 2014; McNeish, 2018). Enligt kritiker av koefficient alfa är det ytterst sällan som dessa förutsättningar är uppfyllda. Andra (t.ex. Ryakov & Marcoulides, 2019) har framhållit att så länge frågorna i ett test är välkonstruerade enligt psykometriska principer, så är avvikelsen mellan

koefficient alfa och teoretiskt möjlig reliabilitet försumbar och bör inte påverka tolkningen av testpoäng. Forskaren bör alltså noga överväga om datasetet i en studie i tillräcklig grad lever upp till förutsättningarna för koefficient alfa. Annars finns en rad föreslagna alternativ och det som oftast framhålls som överlägset koefficient alfa är koefficient omega (McDonald, 1999).

Testpoängens referenspunkt

Detta avsnitt avslutas med en kommentar om vad som utgör referenspunkt för tolkning av testpoäng. Man brukar skilja mellan normrelaterad och kriterierelaterad tolkning, där den förra sätter en testtagares poäng i relation till andra testtagares poäng, medan den senare sätter en testpoäng i relation till något yttre kriterium. Detta yttre kriterium avgränsar vanligen kategorierna "godkända" prestationer från "underkända" prestationer. Inom kategorin skiljer man däremot inte mellan olika prestationer och det är inte av primärt intresse om testet lyckas fånga variation mellan deltagare i samma kategori. Inom forskning som primärt använder korrelationsdesign, exempelvis forskning om språkbegåvning, är det däremot önskvärt att kunna diskriminera mellan individer längs hela färdighetsskalan. För att ett test ska fungera bra i en korrelationsstudie måste det medge en tydlig normrelaterad tolkning. Ordförrådstest (som i denna avhandling föreslås representera generell språkfärdighet) har ofta konstrueras för pedagogiska syften med en kriterierelaterad tolkning i åtanke (t.ex. hur mycket av ordförrådet i en text behärskas av inlärare på en viss nivå). Om dylika test används i forskning om individuella skillnader i förmåga att lära ett L2, så kan en ny validering behöva göras för det nya syftet med testet.

Den förklarande variabeln: språkbegåvning

I detta avsnitt behandlas först olika teorier om språkbegåvning och dess roll för språkinlärning. Därefter görs en genomgång av hur man har försökt mäta språkbegåvning i test. Språkbegåvning har definierats som en relativt stabil egenskap som varierar mellan individer och som inte förändras på ett betydande sätt av träning, samt att den är domänspecifik för språk, d.v.s. skild från generell intelligens (Skehan, 2002). Vidare antas det att språkbegåvning är flerdimensionell, vilket innebär att den består av olika, och sinsemellan inte nödvändigtvis högt korrelerade, delar. Exakt vilka dessa delar är råder det olika uppfattningar om, men en vanlig beskrivning är att språkbegåvning åtminstone utgörs av förmågor att (i) fonologiskt bearbeta språklig input, (ii) analysera språklig struktur och (iii) behålla varaktiga minnen av språkliga element.

Den klassiska teoretiska modellen av språkbegåvning grundar sig på faktoranalys i samband med konstruktionen av The Modern Language Aptitude Test (MLAT, Carroll & Sapon, 1959), det språkbegåvningstest som fått störst användning i forskning och utbildningssammanhang. Carroll identifierade fyra underliggande faktorer som sinsemellan var svagt korrelerade men som

tillsammans visade ett relativt starkt samband med L2-inlärning: Phonetic Coding Ability, Grammatical Sensitivity, Inductive Language Learning Ability och Associative Memory. Testbatteriet MLAT representerade dock inte dessa faktorer lika mycket eftersom man prioriterade kriterievaliditet, d.v.s att testet skulle kunna förutspå inlärarframgång. Att testet skulle representera den underliggande teorin var mindre intressant vid utvecklingen av MLAT eftersom syftet med testet var helt pragmatiskt, nämligen att spara pengar vid språkutbildning. MLAT har senare kritiserats just för att inte vara grundat i teori, samt att enbart se till prediktiv validitet för en viss typ av språkinlärare (unga och motiverade, alla med L1 engelska) i en viss typ av undervisning (intensiv klassrumsundervisning med med den på 1960-talet populära audiolingval metoden).

Två linjer av kritik kan urskiljas som på 1990-talet började ifrågasätta dåvarande språkbegåvningsforskning, som var starkt dominerad av MLAT. Dels Robinson (2001; 2005) som menade att MLAT inte tog hänsyn till inlärning över lång tid i naturlig L2-miljö, samt Skehan (1998, 2002) som menade att MLAT inte tog hänsyn till kognitiva processer som identifierats andraspråksforskningen. Enligt Robinson borde man undersöka språkbegåvning i interaktion med olika sorters språkundervisning och feedback, samt urskilja olika språkbegåvningsprofiler som skulle kunna dra nytta av olika sorters undervisning. Skehan menade att spåkbegåvningstest borde ta hänsyn till kognitiva processer som gör sig olika mycket gällande i olika faser av språkinlärning. Exempelvis menade han att fonologisk bearbetning borde spela större roll i inledningsfasen av en språkutbildning, medan förmåga till språkanalys blir mer aktuellt i ett något senare skede, varefter minne för lexikogrammatiska strukturer blir dominerande i ett längre perspektiv. Delvis under påverkan av Skehans kognitiva modell, som inkluderar fler aspekter av språkbegåvning än i den klassiska modellen, har forskare under det senaste decenniet fokuserat allt mer på arbetsminnets roll för språkinlärning, samt individuella skillnader i implicit (d.v.s. mindre medveten eller omedveten) inlärningsförmåga.

Det finns flera modeller för arbetsminne men den som spelat störst roll i språkvetenskap eller åtminstone i forskning om individuella skillnader är den modell som beskrivs i t.ex. Baddeley (2003). Modellen beskriver arbetsminne som bestående av tre komponenter: den fonologiska loopen (eng: the phonological loop), den visuospatiala avbilden (eng: the visuo-spatial sketchpad och den centrala samordnaren (eng: the central executive). De två förstnämnda komponenterna behåller representationer av ljud respektive bild i korttidsminnet, och den centrala samordnaren kan exempelvis utföra problemlösning, styra uppmärksamhet eller samordna olika input, samt koda information i långtidsminnet. Inom språkbegåvningsforskning är det framför allt den fonologiska loopen som fått uppmärksamhet, då dess funktion dels visat

sig variera mellan individer och dels uppvisa samband med inlärning av bl.a. ordförråd.

Implicit inlärningsförmåga har undersökts i några studier om språkbegåvning (t.ex. Granena, 2016). Implicit inlärning är till skillnad från explicit inlärning en process som sker huvudsakligen omedvetet, t.ex. när uppmärksamhet riktas mot något annat än det som lärs in implicit. Det kan t.ex. vara syntaktiska mönster som lärs in medan en person är fokuserad på språkligt innehåll i stället för form. Vissa samband mellan implicit inlärningsförmåga och L2-färdighet har observerats men denna nya gren av språkbegåvningsforskningen befinner sig ännu i sin linda.

Test av språkbegåvning

Här beskrivs några språkbegåvningstest som i hög grad skiljer sig åt. Dessutom ges en kort beskrivning av test som antas mäta arbetsminne respektive implicit inlärningsförmåga.

The Modern Language Aptitude Test (MLAT) består av fem delar – Number Learning, Phonetic Script, Spelling Clues, Words in Sentences och Paired Associates. MLAT har relativt framgångsrikt lyckats predicera resultat i intensiv, explicit klassrumsförlagd undervisning för vuxna inlärare på grundläggande/intermediär nivå. Korrelationer med färdighetstest i L2 är i storleksordningen 0.50, vilket är högt med tanke på hur många andra (icke-kognitiva) faktorer som också påverkar språkinlärning. MLAT har följts av liknande testbatteri, exempelvis PLAB (för yngre inlärare), DLAB och VORD (refererade i Skehan, 2012). Emellertid har dessa inte överträffat MLAT i prediktiv validitet, och är liksom MLAT otillfredställande förankrade i modern SLA-teori och/eller inte fritt tillgängliga för forskare.

Ett delvis annorlunda sätt att se på språkbegåvning representeras av teorin CANAL-F – Cognitive Ability for Novelty in Acquisition of Language (Foreign) (Grigorenko, Sternberg & Ehrman, 2000), operationaliserat i testet CANAL-FT. Teorin utgår från tanken att språkinlärning kräver förmåga att hantera mångtydighet och ny information (ambiguity och novelty). Testet är dynamiskt (mätning sker vid olika tidpunkter) och innehåller nio delmoment, som bygger på ett konstgjort språk, ursulu. Fem av delarna ges tillsammans i en första testfas, och till skillnad från MLAT finns det stora likheter med verklig språkinlärning. Exempelvis testar första delen förmåga att sluta sig till okända ords betydelse via kontexten (i en engelsk text finns insprängda ursulu-ord, vilkas betydelse ska bestämmas).

LLAMA (Meara, 2005) bygger delvis på MLAT, men det är datoradministrerat och gratis nedladdningsbart via internet. Det bygger även på bilder, symboler och amerikanska indianspråk, vilket gör LLAMA oberoende av testtagarens L1, till skillnad från t.ex. MLAT och CANAL-FT som förutsätter att testtagaren kan engelska, eller att testet översätts. LLAMA tar

totalt cirka 25–30 minuter att genomföra, vilket gör det till ett tidseffektivt alternativ för språkforskare. Testet beskrivs ingående i Meara (2005).

The High-level Language Aptitude Battery (Hi-LAB) (Linck m. fl., 2013) bygger på kognitiva teorier om andraspråksinlärning och innehåller 13 komponenter som representerar vitt skilda förmågor och därför kan ge mer detaljerad information om vilka förmågor som testtagare besitter, s.k. begåvningskomplex eller begåvningsprofiler. Testbatteriet innehåller både arbetsminnestest och test av implicit inlärningsförmåga, vid sidan av mer traditionella komponenter. Till skillnad från sina föregångare avser Hi-LAB att predicera hög slutnivå (ultimate attainment) i naturliga lärmiljöer, och inte som t.ex. MLAT enbart resultat i slutet av en intensivkurs från grundläggande nivå. Eftersom slutnivå infinner sig efter lång tid i L2-miljö har Hi-LAB hittills endast validerats i tvärsnittsstudier, men en longitudinell studie är under genomförande (Linck m. fl., 2013).

Test av arbetsminne kan delas in i de som mäter enkelt respektive de som mäter komplext arbetsminne. Enkelt arbetsminne brukar kallas för fonologiskt korttidsminne och testas vanligen genom att en testdeltagare upprepar nonsensord eller siffror. Denna process ställer inga krav på den exekutiva samordnaren utan anses vara ett rent mått på den fonologiska loopens kapacitet. Om man däremot upprepar siffror baklänges, så behöver den exekutiva samordnaren vända på ordningsföljden samtidigt som den fonologiska loopen behåller sekvensen tillgänglig för analys. Detta är ett test av mer komplext arbetsminne. Båda dessa typer av test är relativt enkla att konstruera och de brukar producera reliabla svarsdata. Korrelation med L2-test har varit ganska låga, i storleksordningen $r = 0.20$ (Linck m. fl., 2014).

Test av implicit inlärningsförmåga bygger ofta på att det finns en dold regelbundenhet i en serie input, som testtagaren inte är medveten om, under det att fokus är inriktat mot något annat (Rebuschat, 2013). Input kan t.ex. vara en serie ord (i ett mycket förenklat exempel: hund, häst, tröja, katt, hund, mössa...) och testtagaren ombeds trycka på en knapp så fort som möjligt varje gång ordet betecknar en inanimat referent. Reaktionstider förväntas då att bli kortare om inläraren märker att det finns ett underliggande mönster (t.ex. djur och klädesplagg kommer i en viss ordning). Liknande test har uppvisat samband med språkinlärningsförmåga men samtidigt har de ofta problem med låg reliabilitet.

Utfallsvariabeln i språkbegåvningsforskning: uppnådd färdighet i L2

Här behandlas det som språkbegåvningstest ska kunna predicera, det vill säga färdighet i ett L2. Jag argumenterar för att ordförrådstest ger ett bra mått på generell språkbehärskning och presenterar några olika sätt att konceptualisera och testa ordförråd. Avsnittet slutar med en diskussion om vilka kognitiva

egenskaper hos en inlärare som eventuellt kan bidra till bättre ordförrådsutveckling.

Ordförråd som proxy för generell språkbehärskning

Sedan 1970-talet har språkfärdighet i ett L2 kommit att beskrivas i termer av kommunikativ kompetens (Hymes, 1972). Denna språksyn innebar att språkanvändning, och inte bara språkkunskaper, kom att stå i fokus för forskares, lärares och språktestares intresse. Några olika modeller över kommunikativ språkfärdighet har föreslagits sedan dess och de mest inflytelserika har varit modellerna som beskrevs i Canale och Swain (1980) och Bachman (1990) – den senare modellen kanske mer känd i svenska språktestkretsar från Bachman och Palmer (1996). Det som modellerna har gemensamt är att de jämställer en komponent som rör språkets form och en annan komponent som rör situerat språk i användning. Den första komponenten antas innehålla exempelvis färdigheter i uttal, morfosyntax, ordförråd, textbindning, medan den andra komponenten innehåller pragmatiska och sociolingvistiska färdigheter. Dessa färdigheter kompletteras i modellerna med någon form av strategisk kunskap för exempelvis hur en språkinlärare kan hantera situationer då språket inte räcker till. Hulstijn (2015) argumenterade för att man i modeller som dessa bör skilja mellan mer centrala och mer perifera delar. Ordförråd, grammatik och fonologi räknade han till kärnan i en språkmodell, bland annat med argumentet att kärnaspekter av språket kan förekomma utan att de perifera delarna fungerar, men de mer perifera delarna kan inte existera utan att de centrala färdigheterna finns. En lång rad studier har även visat att ordförråd är mycket starkt korrelerat med flera andra språkfärdigheter samt hur språkbrukare uppfattas av bedömare med avseende på generell språkbehärskning (t.ex. Hulstijn m. fl., 2012; de Jong m. fl., 2012).

Konceptualisera och testa ordförråd

Ordförråd i ett andraspråk beskrivs ofta i tre dimensioner, relaterade till bredd, djup och flyt (t.ex. Daller et al., 2007; Henriksen, 1999). Den första dimensionen, bredd, avser hur många ord en person känner till. Det andra, djupet, avser hur mycket som är känt om varje ord. Den tredje avser snabbhet i tillgång till det mentala lexikonet under flytande språkanvändning (jfr Gyllstad, 2013). Som flera gånger påpekats (t.ex. Milton, 2009; Schmitt, 2014; Vermeer, 2001) är djupdimensionen vagt definierad och svår att mäta på ett tillförlitligt sätt. Jag kommer huvudsakligen att fokusera på den första dimensionen, ordförrådets storlek, eftersom den kan testas med mer tillförlitliga metoder (Gyllstad, 2013). Detta är i sin tur avgörande för forskning relaterad till individuella skillnader i språklig förmåga.

Beträffande utformningen av ordförrådstest diskuterade Laufer och Goldstein (2004) två överväganden som styr vilket frågeformat man väljer för att konstruera testfrågor. Det första avser vilken information om ordet som

frågan ska elicitera (ordbetydelse eller form). Det andra gäller hur den informationen inhämtas (genom igenkänning eller återkallelse). Laufer och Goldstein visade att de fyra frågetyperna bildar en implikationell svårighetshierarki, så att (från svårt till lätt) formåterkallelse> betydelsesåterkallelse> betydelseigenkänning> formigenkänning. De två första motsvarar i huvudsak översättning från L1 till L2, respektive från L2 till L1. Den tredje frågetypen kräver att respondenterna väljer mellan olika alternativ när de presenteras med en orddefinition (t.ex. Nation & Beglar, 2007) eller en mening med en lucka (t.ex. Bokander, 2016). Slutligen är ett exempel på den fjärde frågetypen Ja/Nej-checklistans testformat (t.ex. Lemhöfer & Broersma, 2012; Meara & Buxton, 1987) där respondenterna anger om de är bekanta med ett presenterat ord eller inte. Beträffande valet av ord som ska ingå i testet så kan testkonstruktören lägga olika vikt vid olika validitetsevidens. Om innehållsvaliditet sätts i centrum är det avgörande att välja orden representativt, exempelvis från de frekvensband som är av intresse. Detta kan vara en prioritet om målet är t.ex. att diagnostisera läsförståelse via texttäckning. Om kriterievaliditet anses viktigare så är testfrågornas diskriminerande förmåga och korrelation med andra variabler av större vikt. Detta är ofta målet i forskning om individuella skillnader i L2-inlärning.

Kognitiva förmågor som gynnar ordförrådsutveckling

Mycket tyder på att det finns åtminstone två helt olika kognitiva förmågor för att utveckla ordförråd i ett andraspråk. Det första är relaterat till att memorera ord avsiktligt (plugga glosor), och det andra är relaterat till en mer omedveten (statistisk, implicit) process för att stärka kunskapen om varje ord genom upprepade möten i språkanvändning. Man kan därför förvänta sig att känslighet för statistiska regelbundenheter och ordfrekvensinformation är en viktig determinant för förvärv av L2-ordförråd, åtminstone efter en första exponering som kan kräva en mer medveten inlärningsinsats (Ellis, 2002). Det finns ännu begränsad forskning om effekter av implicit förmåga för vuxnas ordförrådsutveckling i L2, men det finns gott om bevis från barns L1-utveckling som visar en avgörande roll för statistisk inlärning av ordförråd (Erickson & Thiessen, 2015). Det är därför rimligt att tro att testmetoder som används i forskning om implicit statistiskt lärande potentiellt kan fungera som lämplighetstester för att förutsäga utfall av andraspråksinlärning hos vuxna. Ett potentiellt problem är att många testmetoder inom statistisk inlärningsforskning har utvecklats för experimentella studier, vilket innebär att de kanske inte har de psykometriska egenskaper som krävs för att på ett tillförlitligt sätt kunna upptäcka individuella skillnader. I den experimentella traditionen utgör individuell variation ofta mätfel som ska undvikas så långt som möjligt. Därför kan sådana testmetoder ge låg tillförlitlighet och underskatta korrelationer när de används i språkbegåvningsforskning där individuella skillnader är av största vikt (Cronbach, 1957; Hedge et al., 2018; Siegelman et al., 2017).

Med traditionella språkbegåvningstest (t.ex. MLAT) har korrelationer med ordförråd inte varit imponerande, men som nämnts ovan finns det bara några få språkbegåvningsstudier som har inkluderat ordförrådstest som beroende variabel. Metastudien av Li (2016) fann att fonologisk bearbetningsförmåga gav den högsta korrelationen med ordförråd (r = .38). Intressant nog har rena minnestest tenderat att inte korrelera högt med ordförråd, trots att memoreringsförmåga ofta uppfattas som centralt för att lära in vokabulär. I stället ger korrelationer med fonologisk bearbetning, fonologiskt korttidsminne och implicit inlärning stöd för uppfattningen att ordförvärv i stor utsträckning beror på andra mekanismer än förmåga att plugga glosor. Detta ger vid handen att språkbegåvningstest som avser att fånga upp ordförrådsutveckling måste inkludera test av statistisk och implicit inlärningsförmåga.

Metod

Deltagarna i samtliga studier var universitetsstudenter som rekryterades via lärare eller annan personal vid respektive lärosäte. Deltagarna i studie 3 var ett urval av dem som deltog i studie 2, medan studie 4 använde ett helt annat urval. Totalt bidrog 640 individer med data till studierna i denna avhandling. Rekryteringsförfarandet innebar att det inte var möjligt att ha full kontroll över att alla gjorde testuppgifterna seriöst och fokuserat. Förhoppningen var att en stor mängd data i viss mån ska kompensera för bristande kvalitet i datainsamlingen, vilket verkar ha varit fallet då reliabiliteten i data inte var oväntat låg (vilket annars kan förväntas om testdata grumlas av t.ex. fusk eller omotiverade testtagare).

Datainsamlingen följde etiska riktlinjer för god forskningssed utgivna av Vetenskapsrådet (2017). Inga känsliga personuppgifter samlades in och alla dataset anonymiserades. Datainsamling i studierna gjordes med de fyra LLAMA deltesten (Bokander, 2020; Bokander & Bylund, 2020), ordförrådstestet SweLT (Bokander, 2016), ett C-test (Bokander, 2020), samt en "tänka högt" procedur (i Bokander & Bylund, 2020). Dataanalysen var huvudsakligen kvantitativ och byggde till övervägande del på analys av korrelationer. Undantag är den probabilistiska analysen i Rasch-studierna, samt viss kvalitativ innehållsanalys i Bokander & Bylund (2020).

De individuella studierna

Här sammanfattas innehållet i de fyra publikationer som ligger till grund för avhandlingen. Det testteoretiska kapitlet utgör en bakgrund till de tre empiriska studierna och placeras därför först. Därpå följer de två artiklar som på olika sätt bidrar med validitetsevidens för språkbegåvningstestet LLAMA. Den fjärde studien behandlar utvecklingen av ordförrådstestet SweLT.

Studie 1
Bokander, L. (u.u.). Psychometric assessment. Ingår i: S. Li, P. Hiver, & M. Papi (Red.), *The Routledge Handbook of Second Language Acquisition and Individual Differences*. Routledge.

I det här kapitlet diskuteras psykometriska överväganden i studiet av hur individuella skillnader i kognitiva förmågor, till exempel språklig förmåga eller arbetsminne, påverkar andraspråksförvärv. Psykometri, eller testteori, beskriver procedurer för att utveckla tester och utvärdera deras lämplighet för ett avsett bedömningsändamål. Viktiga frågor som berör testkonstruktion inkluderar itemanalys, poängräkning, tillförlitlighet och giltighet; dessa ämnen utgör huvuddelen av kapitlet och de diskuteras ur perspektiv från klassisk testteori och item-respons-teori. Det är viktigt med medvetenhet om att bristande uppmärksamhet på psykometriska frågeställningar kan påverka kunskapsproduktion negativt inom ämnet andraspråksinlärning. De sista delarna av kapitlet ger exempel på hur testteoretiska problem har behandlats i tidigare forskning om individuella skillnader i andraspråksinlärning och förslag läggs fram för att öka psykometrisk medvetenhet i forskarsamhället.

Studie 2
Bokander, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning, 70* (1), 11–47.

Under det senaste decenniet har språkbegåvningstestet LLAMA (Meara, 2005) kommit att spela en allt viktigare roll som instrument för forskning om individuella skillnader i språkutveckling. Ett potentiellt allvarligt problem som har påpekats av flera forskare är dock att LLAMA ännu inte noggrant har validerats. Vi adresserade detta problem genom att undersöka den interna validiteten för detta testbatteri. Vi samlade in LLAMA-data från 350 deltagare och utvärderade dessa data med hjälp av klassisk itemanalys, Rasch-analys och principalkomponentanalys, följande ett ramverk för bästa praxis inom utbildning och psykologisk testvalidering. Resultaten visar att endast ett av de fyra delproven (LLAMA B) genererade poäng som passar en latent trait-modell med tillräcklig noggrannhet. Detta visar att forskare som använder LLAMA-batteriet måste tolka sina resultat med försiktighet och även att det finns potential att utveckla och förbättra LLAMA ytterligare.

Studie 3
Bokander, L. (2020). Language aptitude and crosslinguistic influence in initial L2 learning. *Journal of the European Second Language Association, 4*(1), 35–44.

Språkinlärningsförmåga och tvärspråklig likhet mellan elevernas första språk (L1) och andra språk (L2) är två faktorer som man vet underlättar framgångsrikt L2-lärande. Dessa fenomen har dock sällan undersökts tillsammans i samma studie. För att adressera detta forskningsgap i andraspråksinlärning genomfördes denna studie med 92 internationella studenter i svenska som L2, med olika L1-bakgrunder. Deltagarna genomförde först ett språkbegåvningstest (LLAMA, Meara, 2005) i början av en 6 veckors L2-kurs på nybörjarnivå. Deras L1-bakgrund kategoriserades i förhållande till målspråket som antingen liknande (germanskt L1) eller avlägset (icke-germanskt L1). I slutet av kursen genomförde deltagarna ett test av uppnådd behärskning av svenska. Regressionsanalyser av testpoäng i svenska, med språkbegåvning och L1-bakgrund som oberoende variabler, visade att tvärspråklig likhet förklarade minst lika stor variation i L2-prestation som språkbegåvning. När man jämför effekterna av språkbegåvning i de två L1-grupperna, befanns språkbegåvning vara viktigare för inlärare med ett typologiskt liknande L1, än för inlärare med ett mer avlägset L1. Dessutom ger resultaten stöd för teoretiska förslag som framkommit inom språkbegåvningsforskning gällande att fonologisk processförmåga kan vara av särskild betydelse i de tidigaste stadierna av L2 - förvärv.

Studie 4
Bokander, L. (2016) SweLT 1.0 – konstruktion och pilottest av ett nytt svenskt frekvensbaserat ordförrådstest. *Nordand, 11*(1), 9–30.

Ett flervalstest av receptivt ordförråd, baserat på information om ordfrekvens, konstruerades och provades ut i en pilotstudie. Orden samplades ur en svensk korpusderiverad basordlista från frekvensnivåerna 2000, 3000, 5000 och 8000 (definierade som frekvensband med 1000 ord vardera). Studiens deltagare utgjordes av 290 personer med svenska som främmande- eller andraspråk. De flesta testfrågor fungerade väl och reliabiliteten var god förutom i 2K-nivån, där en tydlig takeffekt gav låg varians i mätdata. I linje med vad tidigare forskning har visat, följde testresultaten ett implikationellt mönster med distinkt progression i svårighet från lägre till högre nivå och detta förhållande kunde iakttas både på grupp- och individnivå. Deltagarnas färdighetsnivå (GERS), enligt lärarbedömning eller kursplacering, visade signifikant korrelation med poängresultat på ordtestet, dock något lägre än väntat. Slutligen föreslås en modell för hur testpoäng kan användas för kvantifiering av receptivt ordförråd.

Diskussion

Det övergripande syftet med denna avhandling var att undersöka i vilken utsträckning LLAMA och SweLT kan fungera som valida instrument för forskning om språkbegåvningsaspekter på ordinlärning, baserat på premissen att ordförrådets storlek är en praktisk proxy för allmän L2-färdighet. Det här

avsnittet kommer först att diskutera validitetsevidens till stöd (eller inte till stöd) för en forskningsdesign som inkluderar LLAMA och SweLT i studiet av individuella skillnader i L2-förvärv. Därefter diskuteras i korthet praktiska tillämpningar för språkbegåvningstest i ljuset av vad som framkommit i avhandlingen.

Tidigare forskning har funnit positiva korrelationer mellan deltestet LLAMA D och en uppsättning lexikala mått (Granena & Long, 2013), om än med endast marginell signifikans på grund av det lilla urvalet. Om man antar att C-testet i Bokander (2020) i hög grad involverade ordkunskap, stödjer studiens resultat Granena och Longs (2013) fynd. Det verkar således som om LLAMA D eller en liknande testuppgift med förbättrade psykometriska egenskaper, skulle kunna vara en potentiell kandidat att inkludera i språkbegåvningsforskning riktad mot ordförrådsutveckling. LLAMA B, som i testmanualen (Meara, 2005) föreslås som ett ordinlärningstest, predicerade inte någon L2-varians i studie 3 vilket kan verka udda om man antar att C-testet huvudsakligen var ett test av ordkunskap. Detta resultat kan dock bero på att dessa båda deltest (LLAMA B och D) involverar helt olika aspekter av ordinlärning. LLAMA B liknar att plugga ord från en ordlista eller flash-kort. LLAMA D, å andra sidan, verkar utnyttja mer implicit bearbetning som behövs för att bygga ett ordförråd över en tid. Mer forskning behövs för att ta reda på exakt vad LLAMA D mäter och dess relation till förvärv av ordförråd. När det gäller den interna validiteten i LLAMA bekräftade studie 2 (Bokander & Bylund, 2020) problemet med låg reliabilitet i delar av LLAMA och detta gäller särskilt LLAMA D. Man kan dock konstatera att detta deltest har uppvisat signifikanta korrelationer med L2-inlärning i flera studier. Studie 2 visade även att LLAMA D inte verkar vara ett endimensionellt test, vilket gör mått på intern konsistens olämpliga för skattning av dess reliabilitet. Sammantaget visar detta att det är för tidigt att, baserat på nuvarande kunskap, avfärda LLAMA D som opålitligt i andraspråksforskning.

Det påpekas ofta som en särskild fördel med LLAMA att det är '"språkneutralt" och Rogers m. fl. (2017) fann att LLAMA verkar fungera lika bra med deltagare från olika L1så länge de är bekanta med det latinska alfabetet. Paradoxalt nog kan denna funktion ha bidragit till den lägre prediktiva validiteten för LLAMA i jämförelse med MLAT, eftersom, vilket framgår av litteraturöversikten ovan, mycket tyder på att L2-förmåga är kopplad till L1-förmåga. Därför är det möjligt att ett språkbegåvningstest som är helt okänsligt för L1 inte kan fungera särskilt bra. En möjlig inriktning för framtida språkbegåvningsforskning skulle kunna vara att anta ett mer språktypologiskt kontrastivt perspektiv och konstruera tester som är skräddarsydda för deltagarnas L1, snarare än att vara språkneutrala.

Sammantaget visade resultaten från studie 4 att en förfinad version av SweLT, efter revidering av vissa frågor som inte fungerade som förväntat, skulle kunna ha potential att vara ett användbart forskningsinstrument. Tillförlitligheten befanns vara acceptabel och extrapolering till en

kriterievariabel (CEFR-nivå) var möjlig men inte särskilt exakt. Ett par frågor skulle dock behöva besvaras innan detta test används i forskning om individuella skillnader i språkbegåvning. Det första är att det ännu är okänt hur SweLT skulle fungera med inlärare på lägre nivå, eftersom deltagarna i studien främst befann sig på mellan- eller avancerad nivå. En andra punkt när man diskuterar SweLT:s möjliga roll som forskningsinstrument i studier av individuella skillnader är att syftet med testning i denna typ av forskning skiljer sig från de pedagogiska målen bakom frekvensbandade ordförrådstest. Att använda test baserade på frekvensband kan informera pedagoger om till exempel vilken typ av läsning som skulle vara mest lämplig för inlärarna. Forskning om individuella skillnader söker i stället att maximera variationen mellan deltagarna. Det skulle kunna vara så att detta inte görs bäst med ett frekvensbaserat tillvägagångssätt. Ett mer renodlat psykometriskt tillvägagångssätt som syftar till att maximera diskriminering och reliabilitet, skulle kunna utgöra ett bättre alternativ. Ordförrådsteset LexTALE, utvecklat för psykologisk forskning (Lemhöfer & Broersma, 2012), utformades med hjälp av ordfrekvens som en grov indikation på svårighetsgrad, varefter item med bäst diskriminering valdes ut. Detta verkar som ett lovande tillvägagångssätt för framtida förfining av SweLT, om syftet är att undersöka individuella skillnader i språkinlärning.

Den sista inferensnivån i Kanes (2006) valideringsmodell gäller konsekvenser av testanvändning i praktiken. Även om sådana konsekvenser inte är ett centralt ämne i denna avhandling, skulle det förmodligen vara en allvarlig försummelse att inte säga något alls om det. I inledningen till denna text konstaterades att det har framförts kritik av effektiviteten i de språkprogram som erbjuds vuxna invandrare i Sverige. Vanliga teman i kritiken är att språkkurser inte är individuellt anpassade och att det finns en stor variation i utvecklingshastighet även bland elever som har tilldelats en grupp baserat på deras utbildningsbakgrund (Skolinspektionen, 2018). På ytan kan språkbegåvningstest för placeringsbeslut verka som den perfekta lösningen på detta problem. Språklärare skulle få arbeta med grupper som har en liknande nivå och undervisning skulle kunna skräddarsys efter inlärarnas behov. Det finns dock minst tre stora utmaningar för en sådan lösning. För det första skulle användning av språkbegåvningstest för praktiska beslut om placering i utbildning kräva mycket tillförlitliga tester för att göra besluten motiverade. I sitt nuvarande tillstånd verkar LLAMA inte kunna uppfylla sådana krav, vilket tydligt framkom i Bokander och Bylund (2020). För det andra sker SFI-utbildningen på relativt låg nivå, från nybörjare till lägre mellannivå. I Bokander (2020) visade resultatet att språkbegåvning kan vara en mindre tillförlitlig prediktor för L2-inlärning än typologisk närhet till L1, åtminstone på nybörjarnivå. Det skulle i så fall vara mer meningsfullt att placera nybörjare baserat på L1 än på deras språkbegåvning. För det tredje genomfördes studierna i denna avhandling, liksom de flesta studier om språkkunskaper och

andraspråksinlärning i allmänhet, med deltagare som har relativt hög utbildningsnivå. Detta är en känd begränsning för mycket forskning som gjorts inom andraspråksinlärning och andra relaterade discipliner (Andringa & Godfroid, 2019). Utbildningsnivån för deltagare i SFI kan variera enormt och även inom ett och samma klassrum. Forskning om individuella skillnader i språkinlärningsförmåga, t.ex. med LLAMA, skulle behöva utföras med mer representativa urval än vad som hittills varit fallet, för att ta reda på om rön kan generaliseras till L2 inlärare med låg utbildningsbakgrund och typologiskt avlägsna L1. Det finns således många frågor kvar att besvara innan vi kan förespråka genomförande av språkbegåvningstest för placeringsbeslut i svensk L2-utbildning för vuxna.

Slutsatser och framtida forskning

Denna avhandling undersökte metodfrågor som kringgärdar användning av LLAMA i forskning om individuella skillnader i andraspråksinlärning. Vissa studier som baserat sina resultat på LLAMA har flera hundra citat i Google Scholar (t.ex. Abrahamsson & Hyltenstam, 2008; Granena & Long, 2013). Studier som dessa kan komma att få ett betydande inflytande på hur kunskap konstrueras inom språkvetenskapen. Det är då oroväckande att tvingas konstatera att LLAMA-testerna lämnar mycket att önska från en psykometrisk synvinkel. Avhandlingen diskuterade också det hittills outforskade alternativet i språkbegåvningsforskning att representera L2-färdighet med ordförråd, och SweLT (efter ytterligare finputsning) föreslogs som ett alternativ när målspråket är svenska. Om man skulle vilja använda ordförråd som en proxy för L2-färdighet behöver befintliga språkbegåvningstest kompletteras med test av förmågor som kan antas förutsäga ordförrådsutveckling, vilket exempelvis inkluderar test av arbetsminne, fonologiskt korttidsminne, implicit inlärning. Som visas i denna avhandling verkar LLAMA D vara en intressant kandidat, men mer forskning behövs för att utröna eventuella samband mellan vad LLAMA D mäter och hur ordförråd i ett andraspråk utvecklas över tid. När förbättrade test i framtiden föreligger, återstår naturligtvis att ge sig ut i ett stort antal SFI-klassrum och undersöka om det verkligen går att finna effekter av språkbegåvning på uppnådd språklig nivå i svenska. Innan bra test för ändamålet finns att tillgå, vore det slöseri med alla inblandades tid att initiera storskaliga studier.

References

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, *30*(4), 481–509. https://doi.org/10.1017/S027226310808073X

Agebjörn, A. (2021). *Learning of definiteness by Belarusian students of Swedish as a foreign language.* Doctoral dissertation, University of Gothenburg.

Andringa, S., & Godfroid, A. (2019). SLA for all? Reproducing second language acquisition research in non-academic samples. *Language learning, 69*, 5–10. https://doi.org/10.1111/lang.12338

Artieda, G., & Muñoz, C. (2016). The LLAMA tests and the underlying structure of language aptitude at two levels of foreign language proficiency. *Learning and Individual Differences*, *50*(Supplement C), 42–48. https://doi.org/10.1016/j.lindif.2016.06.023

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing* (1st edition). Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. OUP Oxford.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Elsevier. https://doi.org/10.1016/S0079-7421(08)60452-1

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173. https://doi.org/10.1037/0033-295X.105.1.158

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing*, *16*(2), 131–162. https://doi.org/10.1177/026553229901600202

Bokander, L. (2016). SweLT 1.0: Konstruktion och pilottestning av ett nytt svenskt frekvensbaserat ordförrådstest. *Nordand: nordisk tidsskrift for andrespråksforskning*, *11*(1), 39–60.

Bokander, L. (2020). Language Aptitude and Crosslinguistic Influence in Initial L2 Learning. *Journal of the European Second Language Association*, *4*(1), 35. https://doi.org/10.22599/jesla.69

Bokander, L. (forthcoming). Psychometric assessment. To appear in S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge handbook of second language acquisition and individual differences*. Routledge.

Bokander, L., & Bylund, E. (2020). Probing the Internal Validity of the LLAMA Language Aptitude Tests. *Language Learning*, *70*(1), 11–47. https://doi.org/10.1111/lang.12368

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences,* (3rd ed.). Routledge.

Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp - the corpus infrastructure of Språkbanken. Proceedings of LREC 2012, 474–478. Istanbul: ELRA

Buchner, A., & Wippich, W. (2000). On the Reliability of Implicit and Explicit Memory Measures. *Cognitive Psychology*, *40*(3), 227–259. https://doi.org/10.1006/cogp.1999.0731

Buffington, J., Demos, A. P., & Morgan-Short, K. (2021). The reliability and validity of procedural memory assessments used in second language acquisition research. *Studies in Second Language Acquisition* (early online view), 1–28. doi:10.1017/S0272263121000127

Bylund, E., Abrahamsson, N., & Hyltenstam, K. (2010). The Role of Language Aptitude in First Language Attrition: The Case of Pre-pubescent Attriters. *Applied Linguistics*, *31*(3), 443–464. https://doi.org/10.1093/applin/amp059

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *I*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Carroll, J. B. (1958). A factor analysis of two foreign language aptitude batteries. *Journal of General Psychology*, *59*, 3–19. https://doi.org/10.1080/00221309.1958.9710168

Carroll, J. B. (1962). *The prediction of success in intensive foreign language training*. s.n.

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83–118). Newbury House.

Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test* (p. 27). Psychological Corporation.

Chomsky, N. (1965). *Aspects of the theory of syntax*. M.I.T. Press.

Christiansen, M. H. (2019). Implicit Statistical Learning: A Tale of Two Literatures. *Topics in Cognitive Science*, *11*(3), 468–481. https://doi.org/10.1111/tops.12332

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*(3), 356–371. https://doi.org/10.1016/j.cognition.2009.10.009

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? In W. S. Sossin, J., C. Lacaille, V. F. Castellucci, & S. Belleville (Eds.), *Progress in Brain Research* (Vol. 169, pp. 323–338). Elsevier. https://doi.org/10.1016/S0079-6123(07)00020-9

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684. https://doi.org/10.1037/h0043943

Crossley, S.A., Salsbury, T., & McNamara, D.S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*(2), 243-263.

Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). *Modelling and Assessing Vocabulary Knowledge*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667268

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Deng, L., & Chan, W. (2017). Testing the Difference Between Reliability Coefficients Alpha and Omega. *Educational and Psychological Measurement*, *77*(2), 185–203. https://doi.org/10.1177/0013164416658325

Dörnyei, Z., & Ryan, S. (2015). *The Psychology of the Language Learner Revisited* (1 edition). Routledge.

Doughty, C. J. (2013). Optimizing post-critical-period language learning. In Giesla Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude and ultimate L2 attainment* (pp. 153–175). John Benjamins Publishing.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, *37*, 66–108. https://doi.org/10.1016/j.dr.2015.05.002

Forsbom, Eva 2006: Deriving a base vocabulary pool from the Stockholm Umeå Corpus. Hämtad 15122015 från http:// stp.lingfil.uu.se/~evafo/resources/base formmodels/

Gass, S., Loewen, S., & Plonsky, L. (2020). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, 1–14. https://doi.org/10.1017/S0261444819000430

Gass, S., & Plonsky, L. (2020). Introducing the ssla methods forum. *Studies in Second Language Acquisition*, *42*(4), 667–669. https://doi.org/10.1017/S0272263120000364

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*(4), 513–543. https://doi.org/10.1017/S0142716406060383

Granena, G. (2013a). Cognitive aptitudes for second language learning and the LLAMA Language Aptitude Test. In G. Granena & M. Long (Eds.), *Language Learning & Language Teaching* (Vol. 35, pp. 105–130). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.35.04gra

Granena, G. (2013b). Individual Differences in Sequence Learning Ability and Second Language Acquisition in Early Childhood and Adulthood. *Language Learning*, *63*(4), 665–703. https://doi.org/10.1111/lang.12018

Granena, G. (2019). Cognitive aptitudes and l2 speaking proficiency: links between LLAMA and Hi-LAB. *Studies in Second Language Acquisition*, *41*(2), 313–336. https://doi.org/10.1017/S0272263118000256

Granena, G, & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *29*(3), 311–343. https://doi.org/10.1177/0267658312461497

Grigorenko, E. L., Sternberg, R. J., & Ehrman, M. E. (2000). A Theory-based Approach to the Measurement of Foreign Language Learning Ability: The CANAL-F Theory and Test. *Modern Language Journal*, *84*(3), 390–405.

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective – challenges and potential solutions. In *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*. (pp. 11–28).

Gyllstad, H. (2019). Measuring knowledge of multiword items. In S. Webb (Ed.), The Routledge handbook of vocabulary studies, 387–405. Routledge.

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL - International Journal of Applied Linguistics*, *166*(2), 278–306. https://doi.org/10.1075/itl.166.2.04gyl

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Henning, G. (1991). A study of the effects of contextualization and familiarization on responses to the TOEFL vocabulary test items. *ETS Research Report Series*, *1991*(1), 1–16. https://doi.org/10.1002/j.2333-8504.1991.tb01390.x

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303–317. https://doi.org/10.1017/S0272263199002089

Hulstijn, J. H. (2011). Language Proficiency in Native and Nonnative Speakers: An Agenda for Research and Suggestions for Second-Language Assessment. *Language Assessment Quarterly*, *8*(3), 229–249. https://doi.org/10.1080/15434303.2011.565844

Hulstijn, J. H. (2015a). *Language Proficiency in Native and Non-native Speakers: Theory and research*. John Benjamins Publishing Company.

Hulstijn, J. H. (2015b). Explaining phenomena of first and second language acquisition with the constructs of implicit and explicit learning. The virtues and pitfalls of a two-system view. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages*, 25–46. John Benjamins Publishing Company.

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*(2), 203–221. https://doi.org/10.1177/0265532211419826

Hymes, D. (1972). On Communicative Competence. In J. B. Pride, & A. Holmes (Eds.), Sociolinguistics: Selected Readings. Harmondsworth: Penguin: 269–293.

Jong, N. H. de, Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5–34. https://doi.org/10.1017/S0272263111000489

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, *44*(2), 137–166. https://doi.org/10.1017/S0261444810000509

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.) *Educational measurement* (4th edition)m pp. 17–64.. Praeger Publishers. https://eduq.info/xmlui/handle/11515/34503

Kaufman, S. B., Deyoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*(3), 321–340. https://doi.org/10.1016/j.cognition.2010.05.011

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*(1), 47–84. https://doi.org/10.1177/026553229701400104

Krashen, S. D. (1985). *The Input Hypothesis: Issues and Implications*. Addison-Wesley Longman Ltd.

Lambelet, A. (2021). Lexical diversity development in newly arrived parent-child immigrant pairs. Aptitude, age, exposure and anxiety. *Annual Review of Applied Linguistics, 41*, 76–94. doi:10.1017/S0267190521000039

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202–226. https://doi.org/10.1191/0265532204lt277oa

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. https://doi.org/10.1177/026553229901600103

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Li, S. (2016). The construct validity of language aptitude. *Studies in Second Language Acquisition*, *38*(04), 801–842. https://doi.org/10.1017/S027226311500042X

Li, S., & Zhao, H. (2021). The Methodology of the Research on Language Aptitude: A Systematic Review. *Annual Review of Applied Linguistics*, 1–30. https://doi.org/10.1017/S0267190520000136

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Smith, B. K., Bunting, M. F., & Doughty, C. J. (2013). Hi-LAB: A New Measure of Aptitude for High-Level Language Proficiency. *Language Learning*, *63*(3), 530–566. https://doi.org/10.1111/lang.12011

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*(4), 861–883. https://doi.org/10.3758/s13423-013-0565-2

Marsden, E., Mackey, A., & Plonsky, L. D. (2016). The IRIS repository: Advancing research practice and methodology. In *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 1–21). Taylor and Francis. https://nau.pure.elsevier.com/en/publications/the-iris-repository-advancing-research-practice-and-methodology

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates Publishers.

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*(3), 389–411. https://doi.org/10.1177/0265532219898380

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Meara, P. (1980). Vocabulary Acquisition: A Neglected Aspect of Language Learning. *Language Teaching*, *13*(3–4), 221–246. https://doi.org/10.1017/S0261444800008879

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142–154. https://doi.org/10.1177/026553228700400202

Meara, P. M. (2005). *Llama Language Aptitude Tests.* Lognostics.

Meara, P. M., & Rogers, V. (2019). *The LLAMA Tests v3.* Lognostics. https://www.lognostics.co.uk/tools/LLAMA_3/index.htm

Messick, S. (1989). Validity. In *Educational measurement, 3rd ed* (pp. 13–103). American Council on Education.

Milton, J. (2009). Measuring Second Language Vocabulary Acquisition. In *Measuring Second Language Vocabulary Acquisition*. Multilingual Matters. https://www.degruyter.com/document/doi/10.21832/9781847692092/html

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control* (pp. xx, 506). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909

Nation, I. S. P. (2013). *Learning Vocabulary in Another Language* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

Nation, I. S. Paul. (1983). Testing and teaching vocabulary. *Guidelines*, *5*, 12–25.

Nation, I. S. Paul, & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13.

Newton, P. E., & Shaw, S. D. (2014). *Validity in Educational & Psychological Assessment*. SAGE Publications Ltd. https://doi.org/10.4135/9781446288856
Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*(1), 1–32. https://doi.org/10.1016/0010-0285(87)90002-8
Oller, J. W. (1979). *Language tests at school: A pragmatic approach*. Longman.
Parry, T. S., & Stansfield, C. W. (1990). *Language aptitude reconsidered*. Prentice Hall Regents.
Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006
Petersen, C. R., & Al-Haik, A. R. (1976). The Development of the Defense Language Aptitude Battery (Dlab. *Educational and Psychological Measurement*, *36*(2), 369–380. https://doi.org/10.1177/001316447603600216
Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery*. Second Language Testing Foundation.
Plonsky, L. (2013). Study quality in SLA: An Assessment of Designs, Analyses, and Reporting Practices in Quantitative L2 Research. *Studies in Second Language Acquisition*, *35*(4), 655–687. https://doi.org/10.1017/S0272263113000399
Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*(2), 538–553. https://doi.org/10.1111/modl.12335
Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the Validity of Quantitative Measures in Applied Linguistics Research1. *Language Learning*, *65*(S1), 37–75. https://doi.org/10.1111/lang.12112
Qian, D. (1999). Assessing the Roles of Depth and Breadth of Vocabulary Knowledge in Reading Comprehension. *The Canadian Modern Language Review*, *56*(2), 282–308. https://doi.org/10.3138/cmlr.56.2.282
Qian, D. D. (2002). Investigating the Relationship Between Vocabulary Knowledge and Academic Reading Performance: An Assessment Perspective. *Language Learning*, *52*(3), 513–536. https://doi.org/10.1111/1467-9922.00193
Qian, D., & Lin, L. (2019). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies.* Routledge.
Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
Raykov, T., & Marcoulides, G. A. (2019). Thanks Coefficient Alpha, We Still Need You! *Educational and Psychological Measurement*, *79*(1), 200–210. https://doi.org/10.1177/0013164417725127
Rebuschat, P. (2013). Measuring Implicit and Explicit Knowledge in Second Language Research. *Language Learning*, *63*(3), 595–626. https://doi.org/10.1111/lang.12010
Robinson, P. (2001). Individual differences, cognitive abilities, aptitude complexes and learning conditions in second language acquisition. *Second Language Research*, *17*(4), 368–392. https://doi.org/10.1177/026765830101700405
Rogers, V., Meara, P., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association*, *1*(1). https://doi.org/10.22599/jesla.24

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926
Schmitt, N. (2014). Size and Depth of Vocabulary Knowledge: What the Research Shows. *Language Learning*, *64*(4), 913–951. https://doi.org/10.1111/lang.12077
Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109–120. https://doi.org/10.1017/S0261444819000326
Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1177/026553220101800103
Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*, *115*(2), 163–196. https://doi.org/10.1037/0033-2909.115.2.163
Service, E., & Kohonen, V. (1995). Is the relation between phonological memory and foreign language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics*, *16*(2), 155–172. https://doi.org/10.1017/S0142716400007062
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.
Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, *49*(2), 418–432. https://doi.org/10.3758/s13428-016-0719-z
Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford University Press.
Skehan, P. (2002). 4. Theorising and updating aptitude. In P. Robinson (Ed.), *Language Learning & Language Teaching* (Vol. 2, pp. 69–93). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.2.06ske
Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In Gisela Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Bilingual Processing and Acquisition* (pp. 17–40). John Benjamins Publishing Company. https://doi.org/10.1075/bpa.3.02ske
Skehan, P. (2019). Language Aptitude Implicates Language and Cognitive Skills. In: Z. Wen, P. Skehan, A. Biedron, S. Li, and R. L. Sparks (Eds.), *Language Aptitude. Advancing theory, testing, research, and practice* (pp. 56–77). Routledge. https://doi.org/10.4324/9781315122021-4
Skolinspektionen. (2018). *Undervisning i svenska för invandrare*. Dnr 400-2016:6995.
Skolverket. (2020). Antalet elever inom komvux fortsätter att öka. Retrieved from https://www.skolverket.se/skolutveckling/statistik/arkiverade-statistiknyheter/statistik/2020-06-11-antalet-elever-inom-komvux-fortsatter-att-oka
SOU (2013). Svenska för invandrare – valfrihet, flexibilitet och individanpassning. Statens offentliga utredningar 2013:76. Retrieved from https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2013/10/sou-201376/
Sparks, R., & Ganschow, L. (1993). Searching for the Cognitive Locus of Foreign Language Learning Difficulties: Linking First and Second Language Learning. *The Modern Language Journal*, *77*(3), 289–302. https://doi.org/10.1111/j.1540-4781.1993.tb01974.x
Sparks, R. L. (2012). Individual Differences in L2 Learning and Long-Term L1–L2 Relationships. *Language Learning*, *62*(s2), 5–27. https://doi.org/10.1111/j.1467-9922.2012.00704.x

Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2009). Long-term relationships among early first language skills, second language aptitude, second language affect, and later second language proficiency. *Applied Psycholinguistics*, *30*(4), 725–755. https://doi.org/10.1017/S0142716409990099

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, *25*(2), 293–321. https://doi.org/10.1017/S0142716404001146

Stansfield, C. W., & Reed, D. J. (2019). The MLAT at 60 years. In Z. (Edward) Wen, P. Skehan, A. Biedroń, S. Li, & R. L. Sparks (Eds.), *Language Aptitude: Advancing Theory, Testing, Research and Practice* (pp. 15–32). Routledge.

Stewart, J. (2014). Do Multiple-Choice Options Inflate Estimates of Vocabulary Size on the VST? *Language Assessment Quarterly*, *11*(3), 271–282. https://doi.org/10.1080/15434303.2014.922977

Stewart, J., McLean, S., & Kramer, B. (2017). A Response to Holster and Lake Regarding Guessing and the Rasch Model. *Language Assessment Quarterly*, *14*(1), 69–74. https://doi.org/10.1080/15434303.2016.1262377

Stoeckel, T., McLean, S., & Nation, P. (2020). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, 1–23. https://doi.org/10.1017/S027226312000025X

Stoeckel, T., Stewart, J., McLean, S., Ishii, T., Kramer, B., & Matsumoto, Y. (2019). The relationship of four variants of the Vocabulary Size Test to a criterion measure of meaning recall vocabulary knowledge. *System*, *87*, 102161. https://doi.org/10.1016/j.system.2019.102161

Suzuki, Y. (2021). Probing the construct validity of LLAMA_D as a measure of implicit learning aptitude: Incidental instructions, confidence ratings, and reaction time. *Studies in Second Language Acquisition*, First View. https://doi.org/10.1017/S0272263120000704

Suzuki, Y., & DeKeyser, R. (2017). The Interface of Explicit and Implicit Knowledge in a Second Language: Insights From Individual Differences in Cognitive Aptitudes. *Language Learning*, *67*(4), 747–790. https://doi.org/10.1111/lang.12241

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*(2), 217–234.

Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the Acquisition of Vocabulary and Grammar: An Individual Differences Approach. *Language Learning*, *70*(S2), 221–254. https://doi.org/10.1111/lang.12395

Webb, S. (2008). Receptive and productive vocabulary sizes of l2 learners. *Studies in Second Language Acquisition*, *30*(1), 79–95. https://doi.org/10.1017/S0272263108080042

Wen, Z. (Edward), Biedroń, A., & Skehan, P. (2017). Foreign language aptitude theory: Yesterday, today and tomorrow. *Language Teaching*, *50*(1), 1–31. https://doi.org/10.1017/S0261444816000276

Winke, P. (2013). An Investigation Into Second Language Aptitude for Advanced Chinese Language Learning. *The Modern Language Journal*, *97*(1), 109–130. https://doi.org/10.1111/j.1540-4781.2013.01428.x

Young-Scholten, M. (2013). Low-educated immigrants and the social relevance of second language acquisition research. *Second Language Research, 29*(4), 441–454.
Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable Sensitivity. *Studies in Second Language Acquisition*, *27*(4), 567–595.
Zhang, D. (2012). Vocabulary and Grammar Knowledge in Second Language Reading Comprehension: A Structural Equation Modeling Study. *The Modern Language Journal*, *96*(4), 558–575. https://doi.org/10.1111/j.1540-4781.2012.01398.x
Zhang, X., & Lu, X. (2014). A Longitudinal Study of Receptive Vocabulary Breadth Knowledge Growth and Vocabulary Fluency Development. *Applied Linguistics*, *35*(3), 283–304. https://doi.org/10.1093/applin/amt014