Individual Differences and the Ergodicity Problem

Lowie, Wander M.; Verspoor, Marjolijn H.

Published in:
Language Learning

DOI:
10.1111/lang.12324

Publication date:
2019

LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Individual Differences and the Ergodicity Problem

Wander M. Lowie[a] and Marjolijn H. Verspoor[b]

[a]University of Groningen and University of the Free State and [b]University of Groningen and University of Pannonia

Traditional research into individual differences (ID) in second language (L2) learning is based on group studies with the implicit assumption that findings can be generalized to the individual. In this article, we challenge this view. We argue that L2 learners do not form ergodic ensembles and that language learning data lack stability. The data from our experiment show that even highly similar learners in terms of ID show clearly different learning trajectories over time; however, we did find that those who showed the greatest degree of variability gained the most in proficiency. Such findings lead to the view that group studies and individual case studies are complementary. Group studies give us valuable information about the relative weight of individual factors that may play a role in L2 development, but longitudinal case studies are needed to understand the process of individual learners' development.

Keywords individual differences; ergodicity; Complex Dynamic Systems Theory (CDST); longitudinal data; second language

Introduction

Although everyone seems to be able to learn the first language (L1) rather successfully, some people are much more successful than others in acquiring

We are very grateful to the editors of this special issue for their excellent feedback and guidance in the process of accomplishing this issue. We are also very grateful to the anonymous reviewers for their thorough reading and their detailed feedback.

Correspondence concerning this article should be addressed to Wander Lowie, University of Groningen, Department of Applied Linguistics, P.O. Box 716, 9700 AS Groningen, Netherlands. E-mail: w.m.lowie@rug.nl

DOI: 10.1111/lang.12324

a second language (L2) and additional languages. This observation logically leads to a research goal to determine the nature and basis of these individual differences (ID). If we manage to discover the factors that are most important in predicting successful L2 learning, we may reveal relevant implications and applications for L2 research and language pedagogy. It is therefore not surprising that the study of ID is a major subfield of research into L2 acquisition.

The ID that are included in this field of research usually concern psychological factors varying from language learning aptitude and motivation to anxiety. Given the knowledge we have of these individual factors, it is tempting to try and identify the most influential ID among learners, and this is often done successfully in group studies. However, we argue that this exercise does not work when we trace individuals over time because of the ergodicity principle. According to Tarko (2005), the concept of ergodicity can be understood with the following example:

> Suppose you are concerned with determining what the most visited parks in a city are. One idea is to take a momentary snapshot: to see how many people are this moment in park A, how many are in park B, and so on. Another idea is to look at one individual (or few of them) and to follow him for a certain period of time, e.g. a year. Then, you observe how often the individual is going to park A, how often he is going to park B and so on. Thus, you obtain two different results: one statistical analysis over the entire ensemble of people at a certain moment in time, and one statistical analysis for one person over a certain period of time. The first one may not be representative for a longer period of time, while the second one may not be representative for all the people. The idea is that an ensemble is ergodic if the two types of statistics give the same result. Many ensembles, like the human populations, are not ergodic. (Tarko, 2005)

Basically, the ergodic principle states that we cannot generalize group statistics—especially when we deal with human beings—to the individual, and vice versa, unless the group is an ergodic ensemble. That is why we need two lines of research in applied linguistics: group studies and single-case studies.

This article is organized as follows: After giving a brief overview of the ID literature, we show that recent trends tend to see ID as dynamic entities that change over time and may affect development differentially at different times. We then argue that because of dynamic changes over time, no two individuals will develop in exactly the same manner as development takes place in a nonlinear fashion, with phases of high degrees of variability accompanying

rapid development. After elaborating on the ergodic principle, we investigate whether a group or subgroup of individuals who are similar in many respects can be seen as an ergodic ensemble in that their differences in motivation and aptitude can be related to their gains in proficiency over a year. To evaluate the contribution of individual factors in both group and individual case studies, we created a unique corpus of 22 learners traced longitudinally with 23 sequential measures over time.

The Dynamic Nature of ID

Research into ID in L2 acquisition goes back a long time. Discussions about one of the most prominent factors studied, the motivation for L2 learning, started as early as the 1950s (Gardner & Lambert, 1959) and continue to this day (Dörnyei & Ryan, 2015). The goals of this long tradition of research have been to identify and explore the effects of ID in L2 acquisition. One would expect that after more than 50 years of research we would come closer to answering the most burning questions in this field and that we would now have a fairly established picture of the factors that affect L2 acquisition. However, in spite of the creation of some more refined definitions of the underlying constructs and in spite of a host of studies carried out, we may have to conclude that "the more we learn about individual differences, the more complex the field becomes" (Ehrman, Leaver, Lou, & Oxford, 2003, p. 325).

The number of factors that can potentially affect the process of L2 development is seemingly endless. Aptitude and motivation are very frequently considered in L2 studies. Studies concentrating on learning styles, learning strategies, and personality factors (such as introversion versus extraversion) have also seen a long history of research. Recently, relatively unexplored areas have been added, such as emotion (MacIntyre, 2002), holistic individual personality (McAdams & Pals, 2006), narrative identity (Dörnyei & Ryan, 2015), and circadian rhythm (De Bot, 2013). The optimal research design seems to be the inclusion of all factors imaginable in one model, but this is hardly feasible.

The factors identifying ID are complex and difficult to define, and measured effects strongly depend on the operationalizations used. Aptitude tests like the Modern Language Aptitude Test (MLAT; Caroll & Sapon, 1959) and the Pimsleur Language Aptitude Battery (PLAB; Pimsleur, 1966) still go surprisingly strong, but language learning aptitude as defined in the MLAT or the PLAB test battery is very different from the same construct as defined in the LLAMA aptitude test (Meara, 2005) or in the Cognitive Ability for Novelty in Acquisition of Language—Foreign Test (Grigorenko, Sternberg, & Ehrman, 2000). Robinson (2005) analyzed the different components of aptitude and distinguished

abilities, aptitude complexes, task aptitudes, and pragmatic abilities, which are related to different learning contexts. While the older, traditional test batteries focus on abilities, more recent tests tend to include some of the other levels. The different dimension of aptitude was illustrated by Skehan (1989), who already pointed to the fact that successful language learners may not be successful in all of the aptitude dimensions. Even the most up-to-date tests generally concentrate on the cognitive definition of aptitude and do not choose to include actual communicative abilities. The same can be said about different definitions of motivation. Whereas the traditional Attitude/Motivation Test Battery (reprinted in Gardner, 1985) is deeply rooted in social psychology with a major distinction between integrative and instrumental orientation, later tests have redefined integrativeness and resulted in a L2 motivational self-system that works out the individual identification of possible selves (Dörnyei, 2005). In this approach, the learner's ideal self expresses the internal drive of integrativeness, while the ought-to self expresses the perceived obligations induced by the learner's environment. The variety of approaches and definitions for each of the ID is also found for personality, from the development toward the Big Five model (Goldberg, 1992) to application of the Myers Briggs Type Indicator, L2 learning (Leaver, Ehrman, & Shekhtman, 2005), and the more recent New Big Five model (McAdams & Pals, 2006). For learning styles and learning strategies, the number of different approaches is perhaps even more compelling to show the numerous inconsistent and changing definitions of ID. Similar to the other factors, the problem is that binary categories like Field Dependent versus Field Independent are rather artificial classifications that ignore the continuity of human cognition. Reviewing the opaqueness of constructs defining ID, Ehrman et al. (2003) concluded that "what we thought were unitary characteristics, like language aptitude . . . , are really ambiguous composites of multiple factors" (p. 325). In addition to the problem of ambiguous composites of multiple factors, studies often ignore the interaction of these factors and the directionality of the interaction. For example, components of motivation are likely to interact with achievement. It may be difficult to progress without motivation, but motivation may also be boosted by achievement (see, e.g., Gardner, Tremblay, & Masgoret, 1997). Achievement, in turn, may be related to aptitude.

Another major challenge for research into ID is the fact that these factors may not be stable but fluctuate over time. Although some factors (like aptitude) within certain definitions may be assumed to remain relatively stable, other factors are bound to fluctuate strongly. This assumption is supported by the recent emergence of subfields of motivation research as "motivational dynamics" (Dörnyei, MacIntyre, & Henry, 2014). A relevant point in this sense is the time

scale of measurement. Wanninge, Dörnyei, and de Bot (2014) studied motivational dynamics in Spanish classes at a rather short time scale of 5 minutes. They showed that between the 5-minute steps, motivation was highly variable with rather large differences in variability between the individual learners. Apparently, ID that appear to be stable between two long-term measurements may turn out to be variable at a finer time scale, and it is not unlikely that at still shorter time scales embedded patterns of variability may show (see, e.g., De Bot, Chan, Lowie, Plat, & Verspoor, 2012). Dörnyei (2009, 2010) proposes to redefine ID as dynamic properties representing individually motivated change over time. Although this approach acknowledges the fact that ID are not stable over time, it creates a serious challenge for measuring them, as not only will the factors themselves change over time but so will the interaction with other factors. Rather than viewing ID as stable and mutually exclusive categories that interact in a deterministic manner, a more realistic representation is that of constructs that are not mutually exclusive and show complex dynamic interactions over time. And because the dynamic character of the interaction of these fuzzy categories is likely to be individually determined, group measurements of these interactions may not say much about the individual.

Finally, the factors defined in terms of ID tend to interact with numerous other factors like the learning context, age-related factors, other languages the learner may be familiar with, and so on. DeKeyser (2000) showed that dimensions of aptitude are different at different ages. Older learners tend to be more dependent on analytic abilities as measured by aptitude tests than younger learners. Also in regard to aptitude, Robinson (2005) showed that the accuracy of predictions by MLAT depend on levels of proficiency. MLAT was effective in predicting outcomes of language learning at early stages of development but less effective at advanced stages of proficiency. The strong focus on cognitive processing may have often underestimated the dynamic interaction with the social context in which the ID are situated. A minor change in social interaction may lead to subsequent changes in motivation, achievement, and learning style.

The realization that very few variables remain the same over time has led to longitudinal designs. Studies in the context of complex dynamic systems (De Bot, Lowie, & Verspoor, 2007; Larsen-Freeman & Cameron, 2008) have demonstrated the relevance of including multiple measurements in time to capture the changing L2 system, which is hardly ever stable and cannot be characterized as predetermined and linear. Researchers who investigate ID are also becoming more and more aware of the limited stability of ID (Dörnyei, 2009). Not all factors are equally variable, and the variability may depend on the time scale. For instance, language learning aptitude and working memory

may be considered as relatively stable at shorter time scales but do tend to change across the life span (Waters & Caplan, 2003). Motivation, however, has been shown to vary on several time scales, from minutes (Wanninge et al., 2014) to the lifespan (Kormos & Csizér, 2008). The recent rise of motivational dynamics (Dörnyei et al., 2014) shows that ID are also increasingly investigated with multiple measures over time. There are plenty of studies that show how motivation changes over time within the individual. Jiang and DeWaele (2015) showed how the different integrative aspects of motivation, ought-to self, and ideal self change across three measurements in time. The study also showed that these changes are strongly individual. Jiang and DeWaele (2015) concluded:

> The analyses revealed a complex picture of Ideal/Ought-to L2 self, which changed over time and were affected by various motivational variables. Significant changes occurred in Ideal/Ought-to L2 self and their relationship with other motivational factors over the year. The nonlinear changes in Ideal/Ought-to L2 self was consistent with the basic dynamic features of self-concept. (p. 349)

At least for motivation, we can conclude that case studies focusing on change over time can provide relevant and interesting information about the patterns of language development for single individuals.

However, with all these fluctuations in individual data, the question is what these single case studies can say about the group. For example, Hulstijn (2015) asked what individual case studies focusing on change over time can tell us about predicting the factors that shape the system of language learners in general. The response may be disappointing. Molenaar and colleagues (Molenaar, 2015; Molenaar & Campbell, 2009) pointed out that intraindividual analyses of personality and emotional processes cannot be applied to a group and vice versa, because a group of humans is usually not an ergodic ensemble.

Molenaar and Campbell (2009) argued that the classic ergodic theorem requires that the generalization of observations across individuals can only be made under two strict conditions. The first condition is that the population should be homogenous and the very same statistical model that is used to describe the group as a whole should apply to all subjects in the population. In other words, the means and other descriptive statistics describing the data should not vary across individual participants. Only then can the statistical model of the population be applied to an individual participant from that population. To illustrate violations of ergodicity, Molenaar and Campbell (2009) referred to the repeated measurement of a personality test that 22 participants

completed on 90 consecutive days. The questionnaire consisted of 30 items to test on factors representing components of the Big Five personality factors (Neuroticism, Extraversion, Agreeableness, Conscientiousness, and Intellect). The group analyses showed that the questionnaire reliably explained the factors of the Big Five personality components. However, when looking at the 30 repeatedly measured item scores of each of the individual participants, the Big Five personality factors do not reliably explain the correlations between the scores. The factor loadings were substantially different for each of the individual participants in the test, both in terms of the number of factors involved and in how the factors related to the items in the questionnaire. Because the homogeneity condition is violated by the process of measuring personality with this test, Molenaar and Campbell concluded:

> The nominal interindividual (Big Five) structure cannot be generalized to the level of variation within each subject. Consequently, one cannot expect that the correlations between repeatedly measured items scores of an individual subject can be explained by the factors Neuroticism, Extraversion, Agreeableness, Conscientiousness, and Intellect. (2009, p. 115)

The second condition for ergodicity is stationarity. It requires that the data must be stable and that the mean and variance should not change between the measurements. In other words, the statistical parameters like factor loadings should stay the same across all measurements in time. Molenaar, Sinclair, Rovine, Ram, and Corneal (2009) argued that virtually all studies that focus on change over time of psychological characteristics within individuals violate the stationarity condition for ergodicity of the data. They claimed that the combination of individuals into groups is inappropriate for studies of development, as developmental processes are almost always nonstationary and thus nonergodic. They illustrated this point with data from a study that investigated the development of the emotional experience of eight sons and eight stepsons as they interacted with their fathers during 80 interactions over time. For each single participant, a factor analysis was used to identify three factors: Involvement, Anger, and Anxiety. The authors fitted a nonstationary space–state model to single-subject time-series data using a recursive estimator (EFKIS). This EFKIS dynamic factor analysis of the time series showed that high anxiety initially relates to a decrease in the predicted value of Involvement at the next point in time but later to an increase in the predicted value of Involvement at the next point in time. The time-series model thus showed that the relationship

between anxiety and involvement was dynamic as it changed from a negative relationship to a positive relationship about halfway through the time series. Their study clearly showed that due to the violation of this ergodicity condition, interindividual variation cannot be equated with intraindividual variation.

The focus on individual developmental trajectories over time are fully in line with an approach to L2 development based on the Complex Dynamic Systems Theory (CDST; De Bot et al., 2007; Larsen-Freeman & Cameron, 2008). In this framework, (second) language learning is seen as a holistic process in which all internal and external factors involved continuously interact in a dynamic fashion. In L2 development, the number of relevant subsystems comprise the components of the L1 and the L2 but also include the learner's psychological states and the learner's changing environment. In the context of the current discussion, one of the most relevant characteristics of a complex dynamic system is its variable development and variability is seen as a prerequisite of development and therefore a source of information.

> [Variability in behavior is] . . . especially large during periods of rapid development because at that time the learner explores and tries out new strategies or modes of behavior that are not always successful and may therefore alternate with old strategies or modes of behavior. From a more formal perspective, systems have to become "unstable" before they can change. For instance, high intraindividual variability implies that qualitative developmental changes may be taking place. The cause and effect relationship between variability and development is considered to be reciprocal. On the one hand, variability permits flexible and adaptive behavior and is a prerequisite to development. (Just as in evolution theory, there is no selection of new forms if there is no variation.) On the other hand, free exploration of performance generates variability. Trying out new tasks leads to instability of the system and consequently to an increase in variability. Therefore, the claim is that stability and variability are indispensable aspects of human development. (Verspoor & van Dijk, 2013, pp. 651–652, citations omitted)

Because both the number of factors and the interaction of these factors will be different for different learners, the process of L2 development has been characterized as "individually owned" (Lowie, Van Dijk, Chan, & Verspoor, 2017). In view of the highly individual nature of L2 development, the ergodicity argument makes perfect sense. If the individual's development is the result of the complex interaction of different factors at different moments in

time, both of the ergodicity requirements are violated: A randomized group is most probably not homogenous and the data are not stable. Moreover, the variable nature of development also means that the development is not predetermined and not completely predictable. This view on L2 development is therefore incompatible with ideas commonly accepted, such as a fixed order of acquisition regardless of the learner's mother tongue, as suggested by the morpheme order studies and underlying influential ideas like Krashen's Natural Order hypothesis (Krashen, 1981). Although general patterns may be found when focusing on group means, individual data do not support the idea of a fixed order of acquisition (Lowie & Verspoor, 2015). However, Molenaar and Campbell (2009) do suggest that generalization to the wider population can possibly be achieved "through the identification of subsets of similar individuals" (Molenaar & Campbell, 2009, p. 116), which might be considered an ergodic ensemble.

To summarize, we should distinguish two lines of research in L2 acquisition that are complementary. One is the focus on data collected at one moment in time. This is what Molenaar and Campbell (2009) referred to as interindividual data. For this line of research, time is not the issue, and the use of group studies is the most appropriate method as it will allow us to investigate the relative contribution of individual factors that have affected language learning. Especially when using up-to-date statistical techniques, such as mixed-effect modeling with the participant as a random factor included in the model, this type of study can be very informative. Most, if not all, studies concerning ID in L2 learning have taken this approach. The second line of research focuses on development as it evolves over time. For this approach, longitudinal, individual case studies are more appropriate. The goal is to gain insight into the actual developmental process by tracing different subcomponents of the system and plot their behavior and interactions and not to generalize to groups. Various techniques have been developed to evaluate the significance and relevance of observations made in the data such as moving averages, min–max graphs, and moving correlations of change are descriptive techniques used for this purpose (cf. Verspoor, De Bot, & Xu, 2011), evaluate the magnitude and significance of the changes in variability (Van Dijk, Verspoor, & Lowie, 2011), or use modeling techniques in which outcomes predicted by computer simulations based on testing theoretical assumptions are tested against real data (Lowie, Caspi, Van Geert, & Steenbeek, 2011). Other analyses focus on very short time scales analyzing 1/F noise to investigate properties of self-organization (Lowie, Plat, & De Bot, 2014).

The Current Study

The current study investigated the role of ID (motivation and aptitude) in a group study and in 22 longitudinal case studies. Both group analyses and individual variability analyses were applied to evaluate the contribution of motivation and aptitude (while controlling for starting proficiency and exposure to English outside the classroom) to proficiency gains. As the group of learners was quite homogeneous in many respects, the hypothesis was that they can be considered an ergodic ensemble and the findings in the two studies would be congruent.

Participants

The participants in our study were 22 Dutch learners of English who started secondary school at the onset of data collection. These learners were in the same school in a small town in the north of the Netherlands and were of approximately the same age (12–13 years old). The learners had enrolled in an English–Dutch bilingual stream in which about 50% of all classes (from History to Mathematics) were taught through English in a Content and Language Integrated Learning (CLIL) setting. This school setting and the pervasiveness of English in the Dutch environment allow for rather massive exposure to English during the period of observation.

L2 Proficiency Measures

To trace language development in an ecologically valid manner, free written language production data were obtained between November 2015 and May 2016. Every other week, 23 times in total, students were asked to produce short texts on a topic related to their own lives or to topics discussed in class, such as "my first month at school," "Christmas carols," and "the May break." Writing was done digitally on a school computer. Due to incidental absences, most learners had missed two or three writing sessions, leading to a total of 388 writing samples. The texts were scored holistically and analytically.

Each text was scored holistically on English proficiency, operationalized as relative complexity, fluency, and accuracy, as in Verspoor, Schmid, and Xu (2012). The students' writing samples were anonymized and fully randomized for student and sequence of writing. Ten raters were trained until full agreement was reached on a subset of samples on a 5-point scale, with 1 representing the relatively weakest and 5 the relatively strongest piece of writing of the set. After the training session in which the team of raters had created their own benchmarks, each of the remainder of the 388 samples was rated by three raters independently. All samples with more than a one-point difference among the raters were reassessed by two other raters. After this procedure, the rater reliability

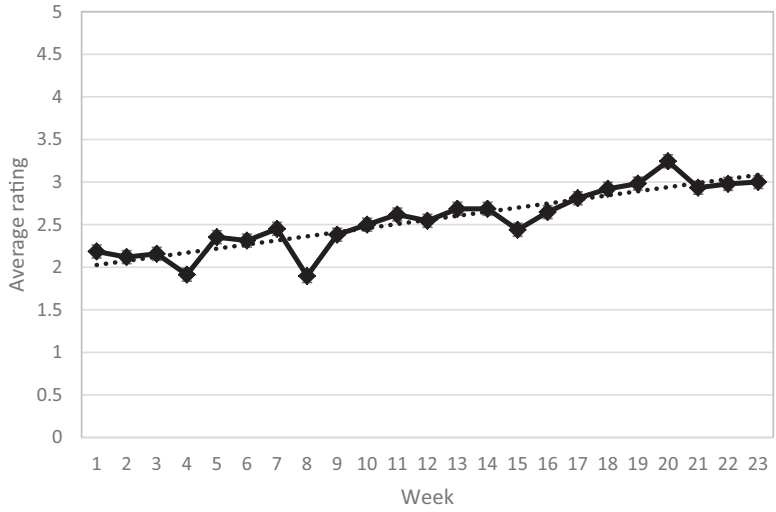

Figure 1 The average rating for each of the weekly topics.

was assessed by calculating an Intraclass Correlation Coefficient (ICC) on absolute agreement (two-way mixed model). The resulting ICC was .78. Then the holistic score for each text was calculated as the average of three ratings.

To control for strong topic effects in text quality, we investigated whether there were clear outliers. For each text (topic), we averaged the score of the group, and these averages were expected to improve over time. Average text length for the topics was 95 words and varied between 82 and 125 words per text, with a gradual increase toward the later samples. Over time, average ratings increased gradually from an average of around 2.1 to 2.9. After correcting for the increasing trend, none of the topics deviated from the expected score, and there was no reason to delete any of the topics from the data set (see Figure 1).

In addition to rating the texts holistically, the writing samples were analyzed on a number of analytical measures that are known to develop over time in learner language (see, e.g., Bulté, 2013). Syntactic complexity was operationalized as the mean length of T-Unit (MLTU) and the number of dependent clauses per T-unit (DC/T). Lexical complexity and lexical sophistication were operationalized as mean length of word (MLW) and Giraud. Complexity of noun phrases was included as the number of dependents per nominal in noun phrases (NomDep). The samples were analyzed using the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (Kyle, 2016; Lu, 2010).

Individual Differences

The students in the group were very similar in many respects because of the selection process. To be allowed into the bilingual stream, students were interviewed and selected on motivation and scholastic aptitude. However, the learners varied somewhat in the number of English classes they had prior to starting at secondary school. To determine more subtle ID between the learners, a number of background variables were obtained through a survey conducted at the beginning of the study. For motivation, three dimensions were included: Ought-to self; Ideal self, and Learning experience, covering Dörnyei and Ushioda's (2009) L2 Motivational Self System. Each component was represented by nine statements, scored on a 6-point Likert scale. For aptitude, the general Dutch Cito score was used, as Verspoor et al. (2011) had shown that even when the bandwidth is relatively narrow, this scholastic aptitude score is a strong predictor of language development. Their group study was done with very similar learners in the Dutch context.

In addition to ID, information on other factors that might contribute to L2 development were obtained, such as the amount of out-of-school exposure to English and the proficiency level of L2 English at the onset of the study, both yielding ordinal scores on a 3-point scale.

Data Analysis

Proficiency gains were operationalized as the difference between early (the average holistic scores of the first two texts) and late proficiency (the average of the last two texts). To relate holistic scores to analytical scores, a regression analysis was run. To see if ID (aptitude and motivation; controlled for starting proficiency and out-of-school exposure) had an effect on proficiency gains, regression analyses were run on these different factors.

To test the hypothesis that strongly similar groups of learners can indeed be considered an ergodic ensemble and show similar effects retrodictively (Molenaar & Campbell, 2009), we compared the individual trajectories of the learners to each other to see if the group of learners developed in similar ways. If the ergodicity of these participants is met, it can be assumed that their developmental trajectories in terms of variability and proficiency gains will be relatively similar.

Because the developmental patterns of the group members were not similar, two subsets of learners were created that were the most similar in as many respects as possible. If the ergodicity of these participants is met, it can be assumed that their developmental trajectories in terms of variability and proficiency gains will be relatively similar.

In line with previous studies (Lowie et al., 2017), we hypothesized that higher proficiency gains coincide with higher degrees of variability. The global amount of variability was evaluated by calculating the coefficient of variation (CV). The CV is a standardized measure of dispersion (*SD*/*Mean*), which is typically stable within individuals. The CV of the individual's holistic scores over time was correlated with the global proficiency gains.

Results

The overall improvement of the learners' writing samples over time was investigated by averaging the ratings of the first two samples of each participant (early) and the last two samples of each participant (late). This general gain is obvious (see Figure 2). This comparison clearly shows a significantly higher average rating of the late samples compared to the early samples, with a large effect (Wilcoxon signed rank = 20.50; $p < .01$; effect size = .84).

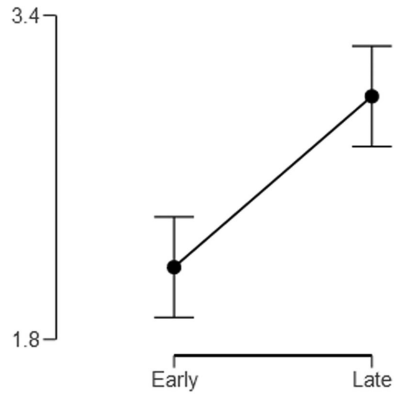

Figure 2 Average first two (Early)—average last two (Late) measurements of the writing samples of the 20 learners.

The improvement over time is illustrated by the writing samples of Student 22 in Week 3 compared to Week 22 of the data collection by the same student, with the early one containing more errors and simpler constructions at sentence, clause, phrase, and word levels.

> Student 22, Week 3: "I like the first week at school the most, because I like playing games and we playing games in the building. First we doing team sports in the Gym building. I like the American Football the most. . . . ."

> Student 22, Week 22: "Vlieland is a wonderful island with friendly inhabitants. Our hotel was at the coast and we came from the harbour to the hotel by a TukTuk (A kind of car). The hotel was nice and there were seagulls everywhere. When we came back to the mainland, I almost fell of the boat (Oops . . . .). This was at the end of the holiday."

To discover which analytical complexity scores predicted the holistic expert ratings best, we performed a regression analysis with Rating as the dependent variable and each of the complexity measures as covariates for all 388 samples. Both MLW and Guiraud were significant predictors for the Ratings (for MLW $\beta = .18$, $p < .001$, for Guiraud $\beta = .47$, $p < .001$). The overall model fit was $R^2 = .27$. None of the other covariates were significant predictors of the expert ratings.

ID as Predictors

A linear regression analysis with the final two ratings (Late) as dependent variable and each of the ID as covariates showed that none of these variables turned out to be a significant predictor of the final ratings. The standardized coefficients of Motivation (.27) and Starting proficiency (.27) were higher than Aptitude (–.04) and Exposure (–.08) but did not reach anything close to significance. The overall model fit was $R^2 = .12$.

Additional linear regression analyses with Motivation, Starting proficiency, Aptitude, and Exposure were run with average scores for each of the complexity measures in the data set (MLTU, MLW, DC/T, NomDep, Guiraud). The results were very similar to those of the holistic ratings. Only with MLW as a dependent variable did we find a trend for the prediction of this variable by the Cito aptitude measure, even though this did not reach significance ($\beta = –.49$; *ns*; overall model fit was $R^2 = .37$).

Individual Trajectories and Ergodic Ensembles

Although the overall ratings of the group are higher for their later writing samples, the variability patterns over time for the individual learners turned out to be quite different. This is illustrated in Figure 3. The four moving average trend lines, which smooth some of the variability, still show different patterns of development for different learners. Correlations of developmental (biweekly) steps between individuals varied (from week to every other week) between $r = .67$ and $r = .06$, with a relatively weak average correlation of .36.

Because the group as a whole did not have similar developmental trajectories, the group as a whole could not be considered an ergodic ensemble.

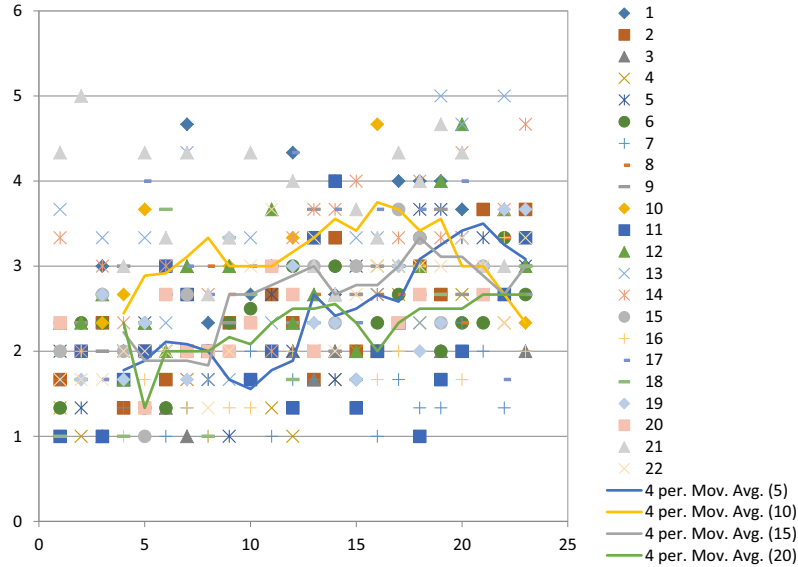

Figure 3 Ratings for all 22 individual participants over time at 23 biweekly writing samples. For learners 5, 10, 15, and 20 moving averages (over four instances) have been added. [Color figure can be viewed at wileyonlinelibrary.com]

Therefore, two subgroups were created that were maximally homogeneous in terms of motivation and aptitude (and other factors such as initial proficiency and out-of-school exposure).

The first subgroup was formed by students who had near-identical high aptitude scores (Cito 548–549 on a 500–550 scale with an observed range of 539–550 in our data), the same high starting proficiency (3), the same high level of exposure (3), and a similar high average of motivation (4.7–4.8). The longitudinal data of this subgroup are shown in Figure 4.

Similarly, the second subgroup was formed by students with identical aptitude scores (Cito 547), the same intermediate starting proficiency (2), a similar low to intermediate level of exposure (1–2), and a similar intermediate average motivation (3.3–3.9), which were some of the lowest scores in the data set. The longitudinal data of this subgroup are shown in Figure 5.

As the figures show, the individuals in each group did not make similar proficiency gains and there was no homogeneity for the groups in this respect (Group1: $M = 1.2$; Group2: $M = 1.3$, with similar dispersion). In both subgroups—which differed in ID—the highest level of achievement was also very similar (Group1: $M = 3.5$; Group2: $M = 3.2$, with identical dispersion).

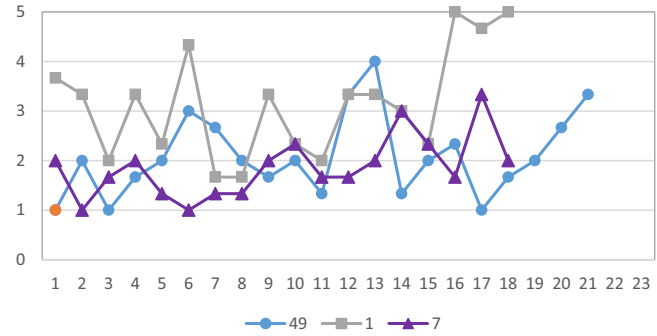

Figure 4 Data of highly similar students (with randomized identification numbers) in subgroup 1. The development represents a continuous sequence—absolute week numbers may be different among participants. [Color figure can be viewed at wileyonlinelibrary.com]

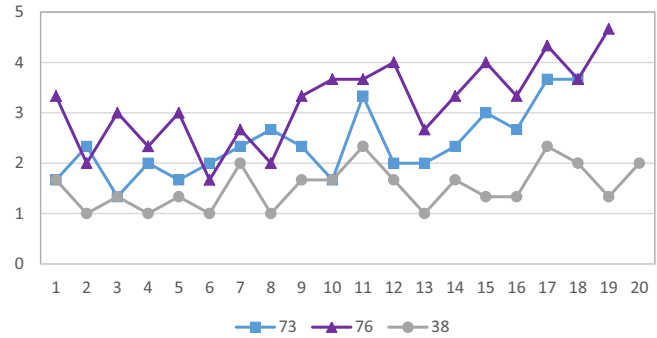

Figure 5 Data of highly similar students (with randomized identification numbers) in subgroup 2. The development represents a continuous sequence—absolute week numbers may be different among participants. [Color figure can be viewed at wileyonlinelibrary.com]

To test for similarities in their developmental trajectories, we examine degrees of variability. For each individual in each group, we calculated the CV of the ratings. Here we do find homogeneity within the groups. In spite of the very small group sizes ($n = 3$), the ratings for Group 1 were significantly more variable (CV = .36, $SD = .03$) than Group 2 (CV = .27, $SD = 0.1$) ($t[4] = 3.5$; $p < .05$; Cohen's $d = 2.8$).

Finally, to explore the relevance of individual variability over time as a meaningful dynamic measure, correlations were calculated between the CV and the global proficiency gains for all participants in the experiment. This turned

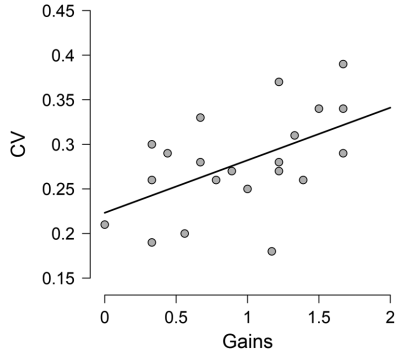

Figure 6 Correlation between coefficient of variance and proficiency gains ($r = .58$).

out to be a moderately strong positive correlation that reached significance ($r_{xy} = .53$; $p < .05$). A higher degree of variability coincided with higher overall proficiency gains (see Figure 6).

Discussion

The present study aimed to show that both group studies and individual case studies are needed in exploring L2 development. Group studies can give us information about different factors such as ID that might play a role in the developmental process. However, the findings may not hold for individuals as they develop over time. Longitudinal case studies, on the other hand, show how an individual develops, but the findings may not hold for other learners. Molenaar and Campbell (2009), however, did suggest that generalization to the wider population can possibly be achieved "through the identification of subsets of similar individuals" (p. 116), which we have called ergodic ensembles. Our assumption was that if we took a very homogeneous group in which as many variables are controlled for as possible in an ecologically valid classroom study, we would find generalizable patterns in the individual trajectories.

In our study, we thus traced the development of English writing of 22 highly similar Dutch learners of English in a semi-immersion CLIL environment. Each learner wrote about 23 texts over the course of one academic year, and each text was rated holistically by experts on relative weak or strong L2 proficiency on a scale from 1 to 5. In addition, each text was analyzed on analytical measures (MLTU, Guiraud, MLW). The observation that the quality of writing increased significantly for the group in this context is not surprising. A regression analysis with Rating as the dependent variable and each of the complexity measures

as covariates for all 388 samples showed that both MLW and Guiraud were significant predictors for the Ratings (for MLW $\beta = .18, p < .001$, for Guiraud $\beta = .47, p < .001$). The overall model fit was $R^2 = .27$. None of the other covariates were significant predictors of the expert ratings. The fact that only lexical measures were strong predictors is in line with findings by Verspoor et al. (2012), which indicated that at the lower levels of proficiency especially lexical measures seem to progress. The fact that there was a negative trend in MLW and aptitude is odd and difficult to explain. However, one possibility comes to mind. As other CDST-inspired research has shown (e.g., Verspoor et al., 2012), learners seem to overdo things until they get it right (hence the variability). For example, our example learner showed longer words in his final text in the sentence "Vlieland is a wonderful island with friendly inhabitants," but his beautifully long word "inhabitants" is quite out of place; a more appropriate form in this context and genre would have been the word "people." In other words, the more advanced learner may not have overused the longer words.

In the group study, we tested whether ID such as motivation or aptitude could predict gains in proficiency controlling for out-of-school exposure and starting level of English. None of these variables predicted the holistic ratings of writing proficiency. This was surprising, as a larger study by Verspoor, de Bot, and Xu (2011) with very similar learners (but in both monolingual and bilingual programs) and similar holistic measures had shown a clear dynamic interplay as proficiency increased between initial proficiency, scholastic aptitude, out-of-school exposure, and motivation/attitude factors. In the first year, scholastic aptitude and initial proficiency were strong predictors. In the third year, scholastic aptitude no longer played a role, but initial proficiency and motivation/attitude did. We assume that in the current study either the restriction of range effect or group size is the reason that we did not find an effect. This small group of learners were highly similar in motivation and aptitude to begin with, as they had undergone entrance requirements for the bilingual program precisely on the two most pregnant variables: aptitude and motivation. Most learners scored close to ceiling for these measures, and the lack of differentiation is not surprising from this perspective. Therefore, even though there were no effects for motivation and aptitude in this particular group, this study does not contradict the relevance of group studies.

The group analysis did not show effects for the ID in proficiency gains, probably because the learners formed such a homogeneous group. Because so many variables had been controlled for, we had hypothesized that the group could be considered an ergodic ensemble and would show similarities in their developmental trajectories. This hypothesis was not confirmed: Figure 3 shows

how each individual showed his/her own trajectory, and the biweekly steps between individuals did not correlate. Because these results did not meet our expectations, we created two subgroups from the larger group that were even more homogeneous in terms of motivation and aptitude. We selected three learners with relatively high motivation and aptitude (controlled for initial proficiency and out-of-school exposure) and three learners with relatively low motivation and aptitude (controlled for initial proficiency and out-of -school exposure). We explored whether they had similar developmental patterns and made similar gains. Even though the group made significant progress in their writing, as did (virtually) all six individual learners, these two highly homogenous subgroups of individual learners did not show a similarity in their pattern of development or in the correlations with growth-related variables. The individual learners were different in virtually all respects: the starting point of their writing, the endpoint, the amount of increase in the ratings of their samples, the pattern of variability, and the amount of variability. Apparently, even in a homogeneous group, interactions among all relevant variables are different for different participants at different moments in time. We assume that these varying factors play a role in the strongly different developmental trajectories.

However, there is one observation that is striking in these data, as displayed in Figures 4 and 5 and confirmed by the analyses involving coefficient of variation. The subgroup of learners with the high scores for aptitude, motivation, and exposure showed relatively more degrees of variability than the subgroup with relatively low to intermediate scores. This finding was corroborated by the significant correlation between the proficiency gains and the coefficient of variation, where high proficiency gains coincided with high values of CV. In other words, there seems to be an interplay between higher motivation, higher aptitude, higher degrees of variability, and greater proficiency gains, but this will have to be investigated further before it can be generalized. The observation is clearly in line with what has been found in other studies observing L2 development over time from a CDST perspective (Lowie, Van Dijk, Chan, & Verspoor, 2017; Verspoor & Van Dijk, 2013). It is true that some of the variability in the text ratings may be due to topic effects and fluctuating motivation, especially for these beginners as their L2 system is still clearly developing. That is what a CDST perspective would predict. At moments of transition, degrees of variability are relatively higher, with significant developmental peaks in some measures (cf. Van Dijk et al., 2011). Variability is meaningful as a required byproduct of the learning process. Without variability, no learning can take place. The data in the current study seem to show that more variability may be a characteristic of a creative learning process, in which new things are tried out

that may go wrong but lead to an exciting process. This may be a finding that could be generalized to other contexts.

The goal of the present study has been to see to what extent we can generalize ID findings from group studies to the individual and vice versa. Unfortunately, we were not able to find a great deal of congruence between the two types of research. However, we may draw several conclusions. Group studies give us valuable information about the relative weight of individual factors that may play a role in L2 development. However, the researcher must keep in mind that the findings may not be representative for a longer period of time and cannot predict much about any individual's behavior at any point in time.

To understand development at the individual level, we need the time dimension. Longitudinal case studies of L2 development are useful methods to provide complementary information about the process of development. For example, in this particular study, we have shown that variability patterns may be worth investigating further. However, the findings from individual cases cannot be generalized to a population of similar learners. In the case of research into ID in L2 development, this implies that differences between individuals cannot and need not be generalized beyond the individual learners we are observing.

Final revised version accepted 8 August 2018

References

Bulté, B. (2013). *The development of complexity in second language acquisition: A dynamic systems approach* (Unpublished doctoral dissertation). University of Brussels, Brussels, Belgium.

Caroll, J. M., & Sapon, S. M. (1959). *Modern language aptitude test*. New York, NY: Psychological Corporation.

De Bot, K. (2013). Circadian rhythms and second language development. *International Journal of Bilingualism*, *19*, 1–14. https://doi.org/10.1177/1367006913489201

De Bot, K., Chan, H., Lowie, W. M., Plat, R., & Verspoor, M. H. (2012). A dynamic perspective on language processing and development. *Dutch Journal of Applied Linguistics*, *1*, 188–218. https://doi.org/10.1075/dujal.1.2.03deb

De Bot, K., Lowie, W. M., & Verspoor, M. H. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*, 7–21. https://doi.org/10.1017/S1366728906002732

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. New York, NY: Routledge. https://doi.org/10.4324/9781410613349

Dörnyei, Z. (2009). Individual differences: Interplay of learner characteristics and learning environment. *Language Learning*, *59*(Suppl. 1), 230–248.

Dörnyei, Z. (2010). The relationship between language aptitude and language learning motivation: Individual differences from a dynamic systems perspective. In E. Macaro (Ed.), *The continuum companion to second language acquisition* (pp. 247–267). London, England: Bloomsbury Academic.

Dörnyei, Z., MacIntyre, P. D., & Henry, A. (Eds.). (2014). *Motivational dynamics in language learning*. Bristol, England: Multilingual Matters.

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited: Second language acquisition research series*. New York, NY: Routledge.

Dörnyei, Z., & Ushioda, E. (2009). *Motivation, language identity and the L2 self*. Bristol, England: Multilingual Matters.

Ehrman, M. E., Leaver, B. L., & Oxford, R. L. (2003). A brief overview of individual differences in second language learning. *System*, *31*, 313–330. https://doi.org/10.1016/S0346-251X(03)00045-9

Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. London, England: Edward Arnold.

Gardner, R. C., & Lambert, W. E. (1959). Motivational variables in second language acquisition. *Canadian Journal of Psychology*, *13*, 266–272. https://doi.org/10.1037/h0083787

Gardner, R. C., Tremblay, P. F., & Masgoret, A. M. (1997). Towards a full model of second language learning: An empirical investigation. *Modern Language Journal*, *81*, 344–362. https://doi.org/10.2307/329310

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*, 26–42. https://doi.org/10.1037/1040-3590.4.1.26

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. *The Modern Language Journal*, *84*, 390–405. https://doi.org/10.1111/0026-7902.00076

Hulstijn, J. H. (2015). Discussion: How different can perspectives on L2 development be? *Language Learning*, *65*, 210–232. https://doi.org/10.1111/lang.12096

Jiang, Y., & Dewaele, J. -M. (2015). What lies bubbling beneath the surface? A longitudinal perspective on fluctuations of ideal and ought-to L2 self among Chinese learners of English. *International Review of Applied Linguistics in Language Teaching*, *53*, 331–354. https://doi.org/10.1515/iral-2015-0015

Kormos, J., & Csizér, K. (2008). Age-related differences in the motivation of learning English as a foreign language: Attitudes, selves, and motivated learning behavior. *Language Learning*, *58*, 327–355. https://doi.org/10.1111/j.1467-9922.2008.00443.x

Krashen, S. D. (1981). *Second language acquisition and second language learning*. London, England: Pergamon Press.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Doctoral dissertation). Retrieved from http://scholarworks.gsu.edu/alesl_diss/35

Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford, England: Oxford University Press.

Leaver, B. L., Ehrman, M. E., & Shekhtman, B. (2005). *Achieving success in second language acquisition*. Cambridge, England: Cambridge University Press.

Lowie, W. M., Caspi, T., Van Geert, P., & Steenbeek, H. (2011). Modeling development and change. In M. H. Verspoor, K. De Bot, & W. Lowie (Eds.), *A dynamic approach to second language development: Methods and techniques* (pp. 22–122). Amsterdam, Netherlands: John Benjamins.

Lowie, W. M., Plat, R., & De Bot, K. (2014). Pink noise in language production: A nonlinear approach to the multilingual lexicon. *Ecological Psychology*, *26*, 216–228. https://doi.org/10.1080/10407413.2014.929479

Lowie, W. M., Van Dijk, M., Chan, H. P., & Verspoor, M. H. (2017). Finding the key to successful L2 learning in groups and individuals. *Journal of Language Teaching and Learning*, *7*, 127–148. https://doi.org/10.14746/ssllt.2017.7.1.7

Lowie, W. M., & Verspoor, M. H. (2015). Variability and variation in second language acquisition orders: A dynamic reevaluation. *Language Learning*, *65*, 63–88. https://doi.org/10.1111/lang.12093

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*, 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

MacIntyre, P. (2002). Motivation, anxiety and emotion in second language acquisition. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 45–69). Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/lllt.2

McAdams, D. P., & Pals, J. L. (2006). A new Big Five: Fundamental principles for an integrative science of personality. *American Psychologist*, *61*, 204–217. https://doi.org/10.1037/0003-066X.61.3.204

Meara, P. (2005). *Llama language aptitude tests: The manual*. Retrieved from http://www.lognostics.co.uk/tools/llama/llama_manual.pdf

Molenaar, P. C. M. (2015). On the relation between person-oriented and subject-specific approaches. *Journal for Person-Oriented Research*, *1*, 34–41. https://doi.org/10.17505/jpor.2015.04

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*, 112–117. https://doi.org/10.1111/j.1467-8721.2009.01619.x

Molenaar, P. C. M., Sinclair, K. O., Rovine, M. J., Ram, N., & Corneal, S. E. (2009). Analyzing developmental processes on an individual level using nonstationary time series modeling. *Developmental Psychology*, *45*, 260–271. https://doi.org/10.1037/a0014170

Pimsleur, P. (1966). *Pimsleur language aptitude battery*. Rockville, MD: Second Language Testing Foundation.

Robinson, P. (2005). Aptitude and second language acquisition. *Annual Review of Applied Linguistics*, *25*, 46–73. https://doi.org/10.1017/S0267190505000036

Skehan, P. (1989). *Individual differences in second-language learning*. London, England: Edward Arnold.

Tarko, V. (2005, December 29). What is ergodicity? Individual behavior and ensembles. *Softpedia News*. Retrieved from https://news.softpedia.com/news/What-is-ergodicity-15686.shtml

Van Dijk, M., Verspoor, M. H., & Lowie, W. M. (2011). Variability and DST. In M. Verspoor, K. De Bot, & W. Lowie (Eds.), *A dynamic approach to second language development: Methods and techniques* (pp. 55–84). Amsterdam, Netherlands: John Benjamins.

Verspoor, M. H., De Bot, K., & Xu, X. (2011). The role of input and scholastic aptitude in second language development. *Toegepaste Taalwetenschap in Artikelen*, *86*, 47–60. https://doi.org/10.1075/ttwia.86.06ver

Verspoor, M. H., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*, 239–263. https://doi.org/10.1016/j.jslw.2012.03.007

Verspoor, M. H., & Van Dijk, M. (2013). Variability in a dynamic systems approach. In C. A. Chapel (Ed.), *The encyclopedia of applied linguistics* (pp. 6051–6059). Oxford, England: Wiley-Blackwell.

Wanninge, F., Dörnyei, Z., & De Bot, K. (2014). Motivational dynamics in language learning: Change, stability, and context. *The Modern Language Journal*, *98*, 704–723. https://doi.org/10.1111/j.1540-4781.2014.12118.x

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, *35*, 550–564. https://doi.org/10.3758/BF03195534