The assessment of language gain in second language learning has long been a focus of interest to teachers, researchers, and program evaluators. Theories of language learning and definitions of language gain have changed in a number of ways over recent decades. This study focuses on how second language research has redefined gain, both substantively in terms of what has been measured, and inferentially, in terms of putative causes underlying individual differences in observed gain. It concludes that there are two main requirements for inferences about language gain to stand up to critical analysis, regardless of the methodology used to assess language abilities. These are that assessments of ability must occur at least twice over the duration of a program, and that all assessment methods must be reliable and valid.

This report will be of interest to those engaged in the assessment of language gain whether from a teaching or program planning perspective.

The other titles in this series are:

1. Language Audits and Industry Restructuring
   Giselle Mawer, 1991

2. Computer-enhanced Language Assessment
   Chris Corbel, 1993

3. Teachers’ Interactive Decision-Making
   David Nunan, 1993

4. Learner Pathways in the Adult Migrant English Program
   Lilli Lipa, 1995

5. Non-language Outcomes in the Adult Migrant English Program
   Elaine Jackson, 1994

6. From Proficiency to Competencies: A Collaborative Approach to Curriculum Innovation
   Youle Bottomley, Jeannette Dalton and Chris Corbel, 1994

7. The Process Syllabus in Action
   Diana Simmons and Sylvia Wheeler, 1995

8. The Computing Practices of Language and Literacy Teachers
   Chris Corbel, 1996

9. Investigating Learner Outcomes for Clients with Special Needs in the Adult Migrant English Program
   Pam McPherson, 1997

10. Current Practice in the Use of Telematics to Support Distance Learners in the Adult Migrant English Program
    Ann Nicholson, 1997

11. Improving Services for Deaf and Hard of Hearing NESB Adults in Australia
    Donovan V. Cresdee, 1997

Steven Ross is Professor of Linguistics in the School of Policy Studies, Kwansei Gakuin University, Japan. He has been a research associate at the National Centre for English Language Teaching and Research, Macquarie University. His current research is in second language acquisition, language assessment, and program evaluation.
Measuring Gain in Language Programs:

Theory and Research

Steven Ross
Measuring Gain in Language Programs: Theory and Research

Published and distributed by the National Centre for English Language Teaching and Research, Macquarie University, Sydney NSW 2109

Ross, S. J. (Stephen John), 1951-. Measuring gain in language programs: theory and research.

ISBN 10 37–5422
I. Second language acquisition. I. National Centre for English Language Teaching and Research (Australia). II. Title. (Series : NCELTR research report ; 12).

418.007

© Macquarie University 1998

Copyright
All rights reserved. No part of this publication may be reproduced or transmitted in any form, or by any means, without the publisher's permission.

This is the twelfth in the NCELTR Research Report Series.

Series Editor: Geoff Brindley

The National Centre for English Language Teaching and Research (NCELTR) is a Commonwealth Government-funded Key Centre of Teaching and Research established at Macquarie University in 1988. The National Centre forms part of the Linguistics discipline within the School of English Linguistics and Media at Macquarie University. NCELTR is funded by the Commonwealth Department of Immigration and Multicultural Affairs.

The publishers have used their best efforts to contact all copyright holders for permission to reproduce artwork and text extracts.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of figures</td>
<td>iv</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td><strong>Chapter 1</strong></td>
<td></td>
</tr>
<tr>
<td>The concept of language gain</td>
<td>3</td>
</tr>
<tr>
<td><strong>Chapter 2</strong></td>
<td></td>
</tr>
<tr>
<td>Criteria for assessing language gain</td>
<td>7</td>
</tr>
<tr>
<td><strong>Chapter 3</strong></td>
<td></td>
</tr>
<tr>
<td>Gain study research designs</td>
<td>13</td>
</tr>
<tr>
<td><strong>Chapter 4</strong></td>
<td></td>
</tr>
<tr>
<td>Language gain research: methods</td>
<td>17</td>
</tr>
<tr>
<td><strong>Chapter 5</strong></td>
<td></td>
</tr>
<tr>
<td>Language gain research: outcomes</td>
<td>21</td>
</tr>
<tr>
<td><strong>Chapter 6</strong></td>
<td></td>
</tr>
<tr>
<td>Causes of individual differences in language gain</td>
<td>27</td>
</tr>
<tr>
<td><strong>Chapter 7</strong></td>
<td></td>
</tr>
<tr>
<td>Language gain measurement criteria</td>
<td>33</td>
</tr>
<tr>
<td>Conclusions</td>
<td>47</td>
</tr>
<tr>
<td>Bibliography</td>
<td>49</td>
</tr>
<tr>
<td>Glossary of acronyms</td>
<td>57</td>
</tr>
</tbody>
</table>
List of figures

Figure 2.1  Standardised score gain
Figure 3.1  Pre- to post-test design
Figure 3.2  Pre- to post-test design with control group
Figure 3.3  Pre- to post-test design with co-variate
Figure 5.1  TOEIC gain as a function of exposure and intensity
Figure 5.2  CSWE gain probabilities
Figure 7.1  Paired t-test
Figure 7.2  Internal consistency
Figure 7.3  Generalisability coefficient
Figure 7.4  Gain score reliability
Figure 7.5  Gain score reliability with pre-post variances
Figure 7.6  Gain dependability with cutting scores
Figure 7.7  Gain score validity
Figure 7.8  Venn diagrams of ANCOVA design
Figure 7.9  Fisher’s test for ANCOVA
Figure 7.10 Hotelling-Williams t-test
Figure 7.11 Fan spread in language gain
Figure 7.12 Multi-wave observation of SLA gain
Figure 7.13 Time-lagged panel design
Introduction

The assessment of language gain has in recent years become the object of considerable interest by parties not usually associated with the academic study of language acquisition. Much of the interest is not motivated by the usual academic focus on defining the bases of human capacity for language learning. Rather, current examinations of language gain are not infrequently tied to the larger context of national language policy — usually in calls for outcomes rationalising public support for language learning programs for adults whose native language is different from that of the surrounding community. Cyclical changes in economics often bring such calls for accountability for language teaching programs. These calls are usually for ‘concrete’ evidence of language learning gain.

This monograph examines often-used criteria for understanding second language acquisition by adults. It reviews recent literature on the nature of adult second language acquisition (SLA) and links SLA issues with research methodologies essential for gathering evidence of language gain. The first half of the review covers how second language learning has been viewed in recent times, and continues with a review of research into the language acquisition of adult migrants — mainly, though not exclusively, in Canada, the USA and Australia. It then links issues of research to methods of analysis important for providing empirical evidence that language learning has taken place.
The assessment of language gain has been a focus of interest and a vexing problem for as long as second languages have been taught (Kelly 1975; Howatt 1984). Through time the conceptual bases by which language learning and language gain have been understood, and how the notion of language gain has been defined, have changed in a myriad of ways. In recent decades there has been a gradual shift in the conceptual basis for defining language gain. In the zenith of behavioural psychology, second language acquisition was assumed to be preceded by the unlearning of linguistic habits instilled through the childhood acquisition of a native language. A second language was concurrently learned through the separation of first language ‘habits’ from a new set of linguistic behaviours, shaped by judicious use of positive and negative reinforcement techniques (Lado 1961). For adult second language learners, the default was assumed to be the automated knowledge of a first language, which was then transferred to the adult learners’ innate hypothesis-generating capacities. A second language morphosyntax and phonology is not therefore acquired in toto, as a first language inevitably is, but is subject to varying degrees of misinterpretation based on how greatly the first and second languages differ in structural ways.

Language typology and distance were assumed to be major causative factors in acquiring a second language. The relative learnability of a second language’s grammatical structure was thought to depend on the degree of contrastive uniqueness, the potential for an erroneous perception of first and second language similarity (faux amis), coupled with the number of structural features of the two languages that would require a second language learner to reduce, expand, or revise grammatical categories. Inherent structural differences might entail a splitting of a single grammatical feature in the first language into two or more distinct categories in the second. This was considered to predict the most difficult second learning task (Stockwell, Bowen and Martin 1965).

Lesser learning tasks were related to the notion of transferability of the first language and the opacity of the second. If there were features in the second language that were not manifested in the second language, these would indicate language learning difficulties of an order less onerous than those requiring a splitting of categories. Conversely, a morphological, syntactic, semantic or phonological feature in the first language which is not manifested in the second was considered a lesser burden for adult learners. Those features that were considered relatively more learnable were based on a reduction of two or more categories in the first language to only one in the second. These were indeed still more cognitively
demanding than direct correspondence between structures in a first-learned language and a second language. Early conceptualisations of gain were thus often framed actually as the loss of first language influences before a gradual gain of the second language system.

The degree of learnability of a second language was thought to depend mainly on the degree of relatedness between the languages. Second languages learned by speakers of typologically distinct languages require considerable learning time and intensive exposure. Typological difference remains as the major touchstone of language learning thought today. Although linguistic theories have progressed and changed considerably in recent decades, the impact of first language on subsequent language learning, whether it is conceived of as a network of automatised neural pathways (Ellis 1990) or as hierarchies of parameterised principles of Universal Grammar (White 1989; Cook 1993; White 1996 cf Epstein, Flynn and Martohardjono 1996; Schachter 1996) where the first language remains the most pervasive, though indirect, influence on the learning of a second language. In this paradigm, ‘gain’ could be construed as a resetting of a first language (L1)-triggered parameter, manifested in the learners’ developing intuition about the impossibility of an utterance such as: *What did Jane destroy a book about?*

Typological differences imply that language gain is mediated by linguistically-defined learning curves determined by the extent of language distance between the native and the target language.

A requisite for operationalising language gain is a definition of language learning. This task has been an onerous one for applied linguists, for there appears to be a phenomenon that separates two kinds of language knowledge. One is the ability for adults in particular to cognitively process language generalisations and factual details without automaticity in applying their knowledge to production in real time. Conversely, there is some evidence that some language generalisations can become automatised without explicit declarative knowledge, which is analogous to the way naive native speakers are thought to know their own language systems intuitively.

The distinction between knowing language intuitively versus ‘knowing about’ language has led to a number of dichotomies describing two overlapping states of human second language knowledge, which has been considered distinct from the implicit knowledge all native speakers are thought to possess. This distinction has been labelled variously: learning versus acquisition (Krashen 1985); automatic versus controlled processing (Bialystok 1982); and declarative versus procedural knowledge (Anderson 1983). The problem these dichotomies present for defining language gain stems from the likelihood that the notion of gain itself would need to be described in a binary manner if the distinction between language learning and language acquisition is a psycholinguistically valid one. We will, in the present sketch of second language gain, assume that learning and acquisition are similar enough to count as the ‘same’ process — although this may prove in the light of SLA research to be merely a convenient fiction.
Another phenomenon that affects language gain relates to the non-linearity of the process. That is, research has suggested that there are apparent stages of language learning through which adult second language learners appear to progress differentially, apparently independent of the influence of their first language (Pienemann 1984; Pienemann and Johnston 1986). Other accounts of this phenomenon have assumed some constellations of affective factors (Gardner 1985) which serve to slow down the process of acquisition leading to learning plateaus — often manifesting language fossilisation of interlingual forms as idiosyncratic grammars (Selinker 1992).

The upshot of arrested language learning phenomena is that they may indicate deeper psychological factors mediating the speed and ultimate attainment of adult second language learning. These factors are not well known, and are subsumed under the general rubric of individual differences (Bley-Vroman 1989; Skehan 1989; 1991). Whether case studies of language learning are involved (Schumann 1978; Schmidt 1983), or cohort-based differences in language learning gain are examined, the impact of individual differences appears to be pervasive. Individual differences are thus the starting point for our understanding of differential gain patterns in adult second language learning.

The focus of this survey henceforth will be on how second language researchers have construed and defined gain, both substantively in terms of what specifically has been gained, and inferentially, in terms of putative causes underlying individual differences in second language gain.
Chapter 2

Criteria for assessing language gain

The first and foremost factor in evaluating a language gain study is the baseline variable. The question to be asked at the outset is ‘gain compared to what?’ This question focuses essentially on the expectation that there has in fact been a discernible gain ordered temporally — that states of language knowledge or performance have ‘improved’ at a later time relative to an earlier time. Without an earlier assessment as a baseline, no direct comparison can in fact be made. Also essential to such an assessment is the comparability of the observations. The most common approach to direct comparison is the use of identical instruments of observation used at two points in time (test and retest). If the baseline observation is some phenomenon related to language learning (for example, self-confidence) but a different aspect of language knowledge at a later time (for example, listening comprehension), the inference that there has in fact been ‘gain’ may well be a faulty one.

Gain is construed most readily by comparing a state of language knowledge or performance at a later point in time with the state of knowledge or performance at an earlier time. Differences in this observation become the basis for inferring that there has been language change or ‘gain’, or plausibly, in the event of prolonged non-use of foreign language, language attrition or ‘loss’ (Lambert and Freed 1982; Maher 1991). Possible causes of an observed gain are inferred after observations of individual differences, with the conditions of common exposure to the second language (usually through instruction) held constant.

The methods by which gain has been assessed vary in complexity and scope. The adequacy of the observation method has a clear impact on the validity of the inference that there has been language gain. Much of the basis for understanding the adequacy of language gain studies is fundamentally related to the soundness of the research design. The structure of the design factors is thus the point of departure in reviewing language gain methods. In this review, three main types of language gain criteria are outlined: Type I (based on standardised scores); Type II (functional performance assessments); and Type III (grammatical features).

Gains in terms of standardised scores

A common approach to operationalising language gain is by comparisons of individuals or cohorts of individuals who are observed to change over time. Here, the approach is often based on observation of an individual’s relative standing in a population of standardised scores. The use of the
standardised score (Brown 1996) permits the inference that an individual has progressed in relative terms to a population of other language learners. Such a comparison does not necessarily support the inference that the individual ‘can do’ certain second language tasks, unless there is ample evidence of criterion score relatedness with such tasks. Figure 2.1 illustrates language gain expressed in terms of standardised scores (Type I, above). The population may be a cumulative record of all adult language learners assessed with the same instrument (examples of this approach are the Test of English as a Foreign Language — TOEFL, and the Test of English for International Communication — TOEIC). Gain for an individual would be construed from a comparison with a scaled score achieved at an earlier point in time (T1) and a score observed at a later point in time (T2). Here again, the notion of gain is essentially the individual’s score in comparison with the standardised population of language learners. Learners ‘gain’ in relative standing to all other learners, rather than in direct acquisition of the language itself.

**Figure 2.1: Standardised score gain**

An advantage of assessing language gain with the use of standardised scores is that all assessment occasions can be related to a common scale, usually arrived at through form equating methodology (Kolen and Brennan 1995). A further advantage is that standardised assessments can be most easily made internally consistent (reliable). The major shortcoming of these assessments is, however, that they typically do not lend themselves to convenient classroom administration, nor do they typically relate to direct ‘real world’ performance interpretations.

This second shortcoming of standardised assessments limits their usefulness to the process of candidate selection when relative (between
subject) decisions are to be made, rather than to absolute assessment of within-subject language gain per se. When standardised assessments are used for program evaluation, they can be taken at best to indicate that the observed gain is in comparison with the cumulative population of learners, and not specifically to the curriculum, or to important criterion tasks that may in fact constitute the de facto standards for employment or settlement needs (Beretta 1986). The major shortcoming is thus in score interpretation, and it is often the responsibility of test score users to locally define the meaning of standardised test scores in relation to program objectives.

**Gains in terms of task performance**

A second operationalisation of language gain emerges from a comparison of task performances at two different times (Type II). Here the gain is inferred when a lack of competence on a performance assessment at an earlier date is replaced by an observation of successful task completion at a later date. The similarity to standardised assessments is obvious here, in that both types of gain are mediated by a temporal difference. The major difference is that in the competence model of gain, the performance may be dichotomously defined as ‘can do’ or ‘cannot do’ rather than as having a score on the underlying set of traits higher than a given percentage of the population. Task comparisons are usually motivated by a functionalist view of language — that language is used for performance tasks and is encoded in the lexico-grammar to achieve these ends (Halliday 1978). An example of a functional task that may reflect language gain could be an assessment of some common task such as:

‘Can read social sight signs’ (Burrows et al 1996).

Assuming that a candidate could not read social sight signs at some point in time prior to instructional intervention or naturalistic acquisition, the inference that language gain had taken place would be supported by an observation of successful reading of such signs at a later point in time.

A more elaborate approach to gain would be based on a lengthy sampling of performance through the use of simulation or interview. Such approaches often lead to problems for the analysis of gain for two main reasons. The first is that unscripted interviews and simulations are very difficult to construct as generic forms to the degree required to make adequately parallel forms (Brindley and Ross, forthcoming). The second is that performances are often evaluated on ordinal scales — usually with global descriptors used to define the characteristics of the performance (cf Ke and Read 1995). Since the assessments tend to be global and on an ordinal scale, inferences about the magnitude of gain are often problematic because there needs to be extensive intervention before observable gain will emerge on the post-intervention assessment (McNaught and McGrath 1996). An example of a rating scale descriptor of proficiency in a pre- and post-intervention assessment might look like:
Pre-test Rating ‘1’ (‘minimum survival skills’) 
Post-test Rating ‘2’ (‘minimum social proficiency’)

Actual incremental gain may be hidden in such comparisons because the observable gain is from one whole category (or rating descriptor) to another. This can easily lead to the incorrect inference that the intervention was inadequate. Some candidates might in fact be at the very threshold of achieving the rating of ‘2’, while others are still far from it. Both types of learners would still be assessed as ‘not 2’ or unsuccessful.

**Gains in terms of grammatical features**

A third kind of language gain (Type III) relates not to tasks or population parameters, but is narrowly defined as specific structural features of the target language. Here, gain might be observed over time as an emergence of a set of particular grammatical features, which may be interpreted as reflecting an order of acquisition (Gass and Selinker 1994; Ellis 1984; Pienemann and Johnston 1986). Once again, gain is inferred after a temporal comparison of features of interlanguage from an earlier sample evolves into a more elaborate set at a later date. This type of ‘gain’ may be the abandonment of first language influences observed at an early point in the acquisition process.

One example would be the acquisition of negation attached to a modal. At an early stage of acquisition, negation may be pre-verbal:

<table>
<thead>
<tr>
<th>(Time 1)</th>
<th>(Time 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘I no work’</td>
<td>‘I don’t work’</td>
</tr>
</tbody>
</table>

Language gain here would be based on the observation that the bound negative morpheme attached to a modal appears later in time, and supplants the previously existing pre-verbal negation.

The notion of language gain at the most general level is a near metaphor for the actual process of language acquisition. Without specific information about what in fact has been ‘gained’, the metaphor fails to capture the psychological reality of second language acquisition. In order for us to understand what in fact has been acquired, it is of utmost importance to specify in as much detail as possible the kinds of language knowledge — instantiated in tasks, tests, interviews, surveys and questionnaires — that form the basis of the definition of language gain.

The review of the extant literature on language reveals a myriad of definitions of language gain. The common factor among them, however, is that they are predicated on the interpretation of some form of analysis of language use. Linguists may find many structural, functional and
pragmatic aspects of language worthy of analysis. Any or all of these may be the object of analytical interest. We need, therefore, to specify what kinds of outcomes are the object of the inference that gain has been made. To this end, the next section examines typical criterion variables that have been identified as having been gained in the process of second language acquisition.
Chapter 3

Gain study research designs

Underlying the notion of language gain is some form of change. The change observed in an individual adult language learner may be construed as being absolute, or as relative to some other group of comparable learners. Essential to the understanding of gain is the way the research itself is designed. This section will outline some of the features of language gain research design.

A basic type of design takes a cross-sectional sample of individual language learners as the starting point for inferring that there has been gain. The cross-sectional approach is based on sampling from a population. This is because the inferences to be made concern the impact of a program (usually of instruction) on the population. The usual approach to this type of design is to pre-test the group with a reliable measurement instrument, provide some form of extensive instruction or language learning opportunity, then post-test the group with the same measurement instrument or a parallel form. If there are statistically significant differences in the pre- to post-test means, the usual inference is that there has been language gain. The reason for gain, however, remains open to interpretation. Researchers into language acquisition often want to provide causal explanations as to why there has been gain, making such 'one-shot' designs (Figure 3.1) unattractive.

Figure 3.1: Pre- to post-test design

Explicit statistical tests of the hypothesis that there has been gain come in a variety of forms, depending on the assumptions underlying the sampling procedure. These issues are dealt with separately later in this monograph.

The single group research design often includes a battery of background factors with which language gain is correlated. While such studies do not provide the basis for clear causal inference, they can provide valuable information when compiled into an aggregate of similar correlational patterns. Cumulatively, single group correlational designs may provide
repeated observations of the same phenomenon. Although rarely performed in language acquisition studies, meta-analysis (Wolf 1986; Hunter and Schmidt 1990) provides the basis for collating and summarising the results of many individual studies, which in the aggregate may provide the basis for what Long (1990a) refers to as ‘a set of laws theory’.

When the object of gain research is to provide evidence that there has been language gain and to provide an account of why there has been gain, a control group design is usually included. The essentials of the control group design are that gain should logically accrue from the impact of a program of instruction or exposure. If such instruction or exposure is withheld, the assumption is that gain should not occur. The instructed group must be compared with the control group before and after the program is implemented. If there is comparable gain for both groups, the program of instruction is assumed not to be the causal reason for the language gain. The control group design (Figure 3.2) assumes in most instances that the two groups are equivalent in background and aptitude at the outset. The main method of ensuring that this assumption will hold is through random selection of group members. Rarely, it must be noted, is truly randomised selection feasible in the context of educational research. A consequence of this limitation is that non-equivalent groups often must be compared.

**Figure 3.2: Pre- to post-test design with control group**

![Figure 3.2: Pre- to post-test design with control group](image)

In lieu of true randomisation, a commonly used approach is to either match the two groups of learners on the pre-test or to adjust for the initial differences statistically with the use of quasi-experimental research design. Usually, an analysis of co-variance (Huitema 1980; Ross 1992a; Lynch 1996) is used. If gain lines run parallel to each other, the assumption is that there is no impact of the program as a causal factor for the gain. If the initially higher group gains at a much faster rate than the initially lower group, the usual inference, based on adjusted post-test scores, is that there is an impact for the program (Figure 3.3). Here, the fact that there has been gain, and an empirical account for why there has been gain, is an attractive aspect of the analysis of co-variance design.
It is sufficient to say that language gain interpretation is most heavily dependent on robust research design. Apparent causal inferences about gain can often be eliminated when other moderating factors are taken into account (Cohen and Cohen 1983). Indeed, apparent gains themselves can be artefactual if the language test instruments are not adequately reliable and valid.
Chapter 4

Language gain research: methods

The second language acquisition literature is rich with accounts of language gain. The methods by which the gain is measured, however, vary considerably in sophistication. The purpose of the following review is to provide an overview of the types of language gain research methods that have been used, and to provide a general picture of the results. The research reviewed here is limited to second language learning by adults in both second and foreign language contexts.

A widely used quasi-experimental research design is the simple pre- to post-test comparison. This design provides the minimal accounting for gain — that there has been language gain. As noted above, the causal factors motivating the gain often cannot be inferred from the simple pre-post design, unless there is a comparison (control) group or groups included, and other controls are exercised over moderating variables. This design is nonetheless a stalwart of gain studies.

An example of a single group study is that by German and Perkins (1984). This study used a residual gain statistic in reading comprehension research in English as a second language to assess individual differences in improvement accruing from learning and training. The subjects were 34 foreign college students with a variety of native languages. A 35-item cloze test (discussed below) was used as the pre-test. A week later, the same students were given the passage again, intact, and were allowed to study it. The test was then readministered.

Lussier, Turner and Desharnais (1993) conducted a study of appropriate standardised instruments for measuring English as a second language (ESL) proficiency within a language contact experience. The purpose of their study was to operationally define the ESL proficiency level of returning high school exchange students according to standardised language proficiency tests. As in the German and Perkins study, a single group was assessed. Here, a range of one to three tests was administered to the students from a group of five standardised language proficiency tests. The aim was to determine the concordance among different standardised tests, not to make an inference about why there had been more gain for some learners and less for others.

An example of a pre-post gain design is found in Williams (1988). Here, a pre-test-post-test treatment-control group design involved comparing 30 beginning ESL subjects. Experimental students received experimental instruction for 15 weeks and the non-equivalent control group received
forced-speech ESL instruction. Both groups were tested for oral proficiency. Regular and percentage gain scores were analysed using independent group t-tests. Experimental students made statistically significantly greater gains than control students in acquisition of vocabulary and comprehension skills and non-statistically significantly greater gains in syntax and verbal expression.

Edwards (1973) conducted a study to determine how effective a communications program was in providing special assistance to students who were experiencing serious difficulty with the English language. 48 students were used in the experimental group at one secondary school and a further 26 students from a different (but parallel) secondary school formed the control group. Most of the students participating were originally migrants to Canada from Hong Kong. Pre-tests and post-tests were given in listening, speaking, reading and writing. A statistical t-test was carried out on the mean gain scores in each skill area to ascertain if there were significant differences between the experimental and control groups. The analysis of co-variance was also applied to the mean gain scores in reading. The results indicated that the program was having a beneficial effect on the development of positive attitudes towards learning among the students concerned. Sound growth in most of the English language skills was also obtained.

Freed (1995) investigated the perceived fluency of study-abroad students by asking native speakers of French to rate the fluency of two groups of students and describe what factors were most important in their rating. Group 1 was comprised of undergraduates who studied abroad for one semester; group 2 were undergraduates who had studied French only in the formal classroom environment. Results showed a slightly larger gain in interview scores and judges' ratings for the students who had studied abroad.

Lapkin, Hart and Swain (1995) evaluated a Canadian interprovincial exchange program to determine (1) what gains are made in French language (L2) proficiency, (2) whether the gains are different for each skill, and (3) whether the level of proficiency before entering the program affects the amount of gain. Their study involved adolescents (N = 119, grades 10–12) all of whom had a background in either core French or second language immersion. The students were given pre- and post-program tests to assess listening comprehension, general L2 proficiency, and reading comprehension. The students were also required to take a self-administered questionnaire before, during, and after the program. Results indicated that students with lower pre-program proficiency achieved greater results from interprovincial exchange. Self-assessment scores increased on post-program questionnaires.

Eichel (1989) reported on the impact of computer-assisted instruction on content course acquisition. Subjects were 38 community college students of varying language backgrounds enrolled in several levels of ESL courses and randomly assigned to experimental and control groups. All students were pre- and post-tested with the same 100-item criterion-referenced test specific to the content of the course work. Eichel's study suggests that
computer-assisted instruction is a boon to ESL and content learning.

In a study of the influence of first language literacy on gain in second language reading, Harwood (1982) investigated why foreign students enrolled in intensive language institutes in the USA exhibit differences in the time it takes them to achieve levels of reading comprehension that will enable them to study successfully at university level. Pre-tests and post-tests were given in English and Spanish to 79 Latin American students. The Spanish tests were the InterAmerican Series Reading Comprehension Test, and a Spanish cloze test. The English tests were the TOEFL, the Michigan English Language Proficiency Test, the Gates-MacGinitie Reading Comprehension Test, and an English cloze test. The test results were then compiled and correlation coefficients computed. The gains made in one semester on the English tests were then calculated for the top Spanish readers and the bottom Spanish readers. The difference in gain scores was significant.

In another correlational study, Palmer (1989) sought to identify useful methods for measuring progress in English language proficiency among Yemeni students. Palmer’s study examined factors critical to progress such as the profiles of students’ backgrounds and their degree of involvement in academic activities. Measures of English language learning progress included gains on the Test of English as a Foreign Language (TOEFL) over two terms; language course grades; and attendance. Factors expected to influence language learning progress were: age; number of dependents; motivation; self-esteem; learning style; and level of enrolment in English learning. Results show a negative correlation between age, number of dependents, and TOEFL gain. There was a positive correlation between motivation and self-esteem, associated with high TOEFL gain rather than low. A negative correlation between enrolment level and TOEFL gain was indicated for those with low gain. Learning style and enrolment level showed a positive correlation for those with high gain. These results suggest that students in the program who succeed in class are highly motivated and able to adjust their learning styles, characteristics that promote academic success.
Crucial to the conceptualisation of gain is the specification of what, exactly, is gained through the processes of second language acquisition. Here we can find that many different outcomes have been used to operationalise, or define, language gain.

One commonly used outcome is the learner’s ability to provide closure to a text or discourse by supplying missing elements. Such cloze tests are thought be ‘integrative’ (Oller 1979), since they require the language learner to integrate many separate bits of linguistic and pragmatic knowledge into the process of closure. Language assessment using some form of cloze procedure is commonplace in the literature on language gain.

A few examples illustrate the use of cloze. Using conventional multiple-choice cloze tests, Edwards, Wesche, Krashen, Clement and Kruidenier (1984) found that Canadian sheltered content instruction class students had made gains in second language acquisition.

German and Perkins (1984) used the residual gain statistic in reading comprehension research in English as a second language to assess individual differences in improvement accruing from learning and training. Their subjects were 34 foreign college students with a variety of native languages. A 35-item cloze test, with every eighth word eliminated but with the first and last sentences unmutilated, was administered in two linguistics class sections. A week later, students were given the passage intact and allowed to study it. The test was then readministered. A standard reading comprehension test was used as a criterion measure. German and Perkins make inferences about the gain in cloze scores.

The sensitivity of cloze tests to detect valid language gain has been the object of considerable research. Perkins and Hunsaker (1990), for instance, performed an item analysis of cloze items. The effectiveness of various indices in measuring information gain sensitivity in cloze tests is evaluated in their research. It appears that cloze not only shows potential for reflecting gain in systemic language ability, but also in schematic understanding of a domain of content knowledge.

The most tangible skill in second language acquisition is associated with the ability to comprehend and speak the second language. A wide variety of speaking tasks have been used to define this skill, however, which often makes direct comparison difficult. The most common assessment is some form of interview, in which a candidate is asked questions in the second
language, to which he or she responds. The responses may be recorded or may be assessed during the interview by an auditor, or by the interviewer. Here, gain may be inferred from differences in speaking fluency, accuracy, appropriateness, or complexity. The usual basis for inferring gain is the observed difference between a holistic rating of proficiency before and after the instructional or naturalistic exposure to the second language.

Ginsberg (1992) studied the predictors of language gain on various criteria, with particular focus on gender, knowledge of other languages, other individual attributes and characteristics of previous language learning careers, the Modern Language Aptitude Test, grammar-based qualifying exams, and proficiency in other language modalities. Interrelationships among gains in different modalities were also examined. Data presented related to 658 students who studied in four-month overseas programs. The outcome measure in this study was the American Council for Foreign Languages (ACTFL) Oral Proficiency Interview (OPI).

Arter (1984) examined the impact of early SLA exposure and instruction for refugees. The refugees were given oral English proficiency tests upon arrival and again after six months of residence. Demographic information, including data on English language training and employment, was also gathered to examine the effects of these factors on the rate of English language learning. They concluded that English language training promotes language learning more than employment does, regardless of the background or experience of the refugees. The amount of previous education was found to be the most important predictor of skill in English, and age was the most important predictor of gain in skill.

Milleret (1991) discusses the use of the ACTFL OPI for assessing the effect of study abroad on foreign language learning, and provides a detailed account of the use of the Portuguese Speaking Test as a means of assessing participants of a summer program in Brazil. Here again, the focus is on gain of oral proficiency in a foreign language.

Mooney (1991) examined significant predictors of achievement in oral English among Spanish-speaking adults learning English as a second language (ESL). It was hypothesised that the Learning Differences Model made up of the variables of instrumental motivation, age, educational attainment (years of previous schooling), exposure to English and informal use of English would account for significant amounts of variation among ESL tests scores. An oral interview was used to assess proficiency of 100 Limited English Proficient (LEP) adults at 0 hours, 50 hours, and 100 hours of instruction. Data, which was collected during an intake interview, consisted of information regarding student demographics and educational background. Information regarding instrumental motivation (motivation characterised by students’ desire to learn a second language for instrumental reasons such as to obtain or upgrade employment) and informal use of English was obtained through the use of the instrument. The learner variables were examined to detect relationships between them and language performance over time. These factors were also studied regarding their capacity to predict achievement at three intervals: (1) Gain 1 (between 0 and 50 hours) (2) Gain 2 (between 50 and 100 hours) and (3)
Gain 3 (between 0 and 100 hours). Four additional variables (gender, reading and writing ability in Spanish, and employment status) were also examined to detect relationships to language performance and achievement. Mooney found that all five factors together were able to predict a significant amount of variation.

Williams (1988) used a pre-test-post-test treatment-control group design which involved comparing 30 beginning ESL subjects. Experimental students received the treatment for 15 weeks and the non-equivalent control group received a forced-speech ESL instruction regimen. Both groups were tested with the IDEA Oral Language Proficiency Test II. Regular and percentage gain scores were analysed using independent group t-tests. Williams found that experimental students made statistically significantly greater gains than control students in acquisition of vocabulary and comprehension skills and non-statistically significantly greater gains in syntax and verbal expression. A moderately positive correlation was found between students’ positive attitudes toward the delayed speech interactive approach and their gains in achievement.

Freed (1995) explored the notion of gain in speaking fluency. She examined the qualities of perceived fluency by asking native speakers of French to rate the fluency of two groups of students and describe what factors were most important in their rating. Group 1 comprised undergraduates who studied abroad for one semester, group 2 were undergraduates who had studied French only in the formal classroom environment (N = 15 each). All students were given the ACTFL/ILR Oral Proficiency Interview (OPI) pre- and post-semester. Native-speaker judges rated a three-minute segment of each student’s OPI test for fluency.

Language gain researchers who use oral proficiency interview data as the basis for the measure of gain do so in order to describe a clearly recognisable language learning outcome to stakeholders. They often do so, however, at a price. There have been serious questions raised about how much the discourse of oral interviews resembles the use of the second language outside the interview context (Lantolf and Frawley 1985, 1988; Bachman and Savignon 1986; Ross 1992b; Ross and Berwick 1992). Furthermore, there can be serious problems with the reliability of oral interviews unless great investments are made in interviewer and rater training and norming.

A more usual approach in measuring language gain is the use of standardised tests. The meaning of ‘standardised’ appears to vary across test makers and users. Here, two definitions are used. One meaning of a standardised test is that there are multiple forms that sample from the same domain of language knowledge. The purpose of the multiple form approach is to promote comparability across test venues over time. Such standardised tests may in fact be equated statistically so as to ensure comparability and reliable indices of gain. When such multi-form tests are equated, there is standardisation in another sense. Changes in observed scores over time are referenced to a population norm. In such instances, we can understand language ‘gain’ as a change in relative ranking in the whole population of test takers, rather than direct gain of the language.
Given the immense cost in time, expertise and money invested in test development, it is no surprise that many language gain research projects involve the use of standardised tests. Spolsky (1990; 1995) explores the origin of the most widely used instrument, the TOEFL, and provides an understanding of how development of the most commonly used test of academic English came to be standardised over some 35 years of worldwide use. Despite the common use of standardised tests to assess gains, however, a number of authors have questioned the validity of the inferences that can reasonably be made about observed changes in scores. Beretta (1986), for instance, argues that the use of standardised test results as the sole criterion for evaluating language programs is bound to lead to inferential error. Des Brisay and Ready (1991) also claim that an intensive English language program in Indonesia, evaluated with standardised instruments, was not credited with other kinds of worthwhile gain. Their study investigated whether: (1) realistic estimates of training needs were being made; (2) data from other tests would enable better prediction of student outcomes, and therefore of student admission criteria; and (3) some guidelines could be established for balancing test preparation and post-course language use in classroom instruction. They note that the outcomes of their study have implications for both the design of the training program and for the use of gain scores in program evaluation.

Lussier, Turner and Desharnais (1993) conducted a study of appropriate standardised instruments for measuring ESL proficiency within a language contact experience. Their purpose was to operationally define the ESL proficiency level of returning high school exchange students according to standardised language proficiency tests. The English proficiency level of native-French-speaking tenth and eleventh grade students from Quebec (n = 185, aged 15–17) who spent three months in an English-speaking milieu was analysed. A range of one to three tests was administered to the students from a group of five standardised language proficiency tests: SLEP, CELT, MTELP, TOEFL and CAT. Lussier, Turner, and Desharnais found that SLEP might be a test that is below the appropriate level for measuring the ESL proficiency of returning exchange students. They also determined that MTELP was inappropriate because students found the content to be uninteresting and tedious. However, results showed that CELT, TOEFL and CAT tests were all strong possibilities for accurately measuring the ESL proficiency level of returning exchange students. They also determined that MTELP was inappropriate because students found the content to be uninteresting and tedious. However, results showed that CELT, TOEFL and CAT tests were all strong possibilities for accurately measuring the ESL proficiency level of returning exchange students. Taking into account the factors of test components, practical constraints, and the number of equated forms of the test that were available, it was decided that TOEFL should be used in the future as the instrument for pre/post-testing to examine ESL proficiency gain.

In a study with university-level adults, Harville (1990) used the TOEFL as a measuring instrument of English language proficiency. The ESL program designed for this study was competency-based and was implemented in a total immersion approach to instruction. The TOEFL exam was used as a pre- and post-test to measure the gain in English language proficiency. Hypothesis testing was used (1) to compare the gain scores of the pre- and post-TOEFL exam scores of students enrolled in the intensive ESL program; (2) to compare the gain scores of the pre- and post-TOEFL exam
scores of students enrolled in the remedial English program; and (3) to compare the TOEFL exam gain scores of students enrolled in the intensive ESL program and TOEFL exam gain scores of students enrolled in the remedial English program.

Harville’s findings were: (1) there were significant gains in English language proficiency on the part of the students enrolled in the ESL program; and (2) there were no significant gains in English language proficiency made on the part of students enrolled in the remedial English course. Thus, the specifically designed ESL program was successful in terms of gain in TOEFL exam scores.

Weissberg and Stuve (1979) conducted a study to examine the effectiveness of intensive English language training for students entering a university ESL program at various levels of proficiency, and to determine if students’ gains in proficiency could be predicted from initial scores on a standardised proficiency test. The Michigan test was administered twice to a group of Latin American students (N = 63), once upon entering the intensive program, and again after ten months of language training. The sample was divided into three proficiency levels according to students’ initial scores. ‘Improvement’ scores were obtained for each level, and differences in gain rates were calculated within and between groups. A significant improvement in proficiencies for each level was found, although no significant differences were found among the levels. Their results contradict other studies that have suggested a higher gain rate can be predicted for students with initially higher proficiencies.

When different standardised tests are used, the obvious question to be answered is the comparability of the different measurements used. Geranpayeh (1994) conducted a correlational study on the International English Language Testing Service (IELTS) in comparison with the TOEFL. The test scores of 216 Iranian graduate students who took the TOEFL and IELTS were compared. The study found high to moderate correlations between TOEFL and IELTS scores. Comparisons indicated that a score of 6 on IELTS is roughly equated with 600 on TOEFL, which is the minimum requirement for non-native speakers to get admittance to most English-language graduate schools. The scores of the most proficient subjects on the two tests were found to be less comparable than the scores of less proficient subjects. Geranpayeh’s study suggests that language gain cannot be accurately inferred by linear equating of scores on different measurement instruments.

In one of the most detailed language test comparability studies to date, Bachman, Davidson, Ryan and Choi (1995) examined the correlation and content co-variance between the First Certificate in English and the TOEFL. Their results indicated strong correlations between the test scores, but that different facets of language knowledge were being tested across the two batteries. No evidence of language gain was included, however. This research suggests that language gain analysis cannot be adequately conducted if the pre-test is done with one set of measurement instruments, and the post-testing is done with another.
The review thus far has focused on language gain as measured in a variety of research designs. There are, however, other criteria important to the language acquisition and acculturation process that do not directly involve language gain directly. These ‘non-language’ outcomes, while not literally ‘gained’ in the sense we have examined in the studies reviewed above, are nevertheless assumed to develop as a consequence of successful experiences in the second language milieu. The inference that these outcomes evolve during the language learning process is not directly amenable to the analyses we have examined in this review because there is no baseline or starting point, and therefore no basis for comparison. The methodology of non-language outcomes is primarily based on questionnaires and self-assessments given to teachers or learners in a post-hoc manner.

Jackson (1994) examined non-language outcomes through the use of teacher surveys. This approach, while not technically reflecting measured ‘gain’ of the outcomes, indicates that teachers in adult migrant programs have a high degree of confidence that learners benefit in substantive ways from the tuition provided. This, as a general form of teacher self-assessment, may have some validity, but clearly further empirical research is needed.

Non-language outcome studies can be examined with the same criteria as those used to assess the impact of language instruction programs. That is, if there are desirable outcomes, the outcomes can be thought of as consequences of the intervention provided by the program. Providing evidence of this, however, is constrained by the same factors as those affecting language gain — essentially providing evidence that the positive outcomes are value-added (Bryk and Weisberg 1976; Lynch 1996), and are causally related to program impact or intervention. In many instances, without clear differences in measured outcomes, language gain or otherwise, reference to positive outcomes may appear to program reviewers as attempts to rationalise marginally successful instructional programs with whatever positive outcomes can be identified. Fajonyomi and Ala-Adeyemi (1993) for instance, ascribe positive outcomes to bilingual instruction in spite of there being no overall difference between instructional groups in an impact study of Nigerian adult literacy learners. In a study of how self-esteem affects language instruction outcomes, Anderson (1982) presents insights into the self-perceptions of students and the teachers’ observations regarding their students’ abilities in English as a second language. Anderson’s results show that students and teachers do not view the students’ language ability similarly, and that self-esteem may be a factor in motivating students. Here, the issue is not that there is direct evidence that self-esteem is a causal variable in ultimate success, but that it is a nice quality to have. It may be, however, less apparent to language teachers than it is to individual adult language learners themselves.

Studies of gain other than those related to language open up a different (but, arguably, important) area of research. Underlying both language gain and desirable non-language outcomes is the issue of causation. Why do individuals differ in their rate of gain, self-esteem, motivation, or degree of socialisation into a new community? Putative reasons for differences in gains of all kinds are thought to be found in the myriad of possible psychological ways that persons differ from one another.
Chapter 6

Causes of individual differences in language gain

In reviews of the factors affecting individual differences, Skehan (1991) and, more recently, Ehrman (1996) have identified cognitive learning styles, maturational constraints, field independence and field sensitivity, personality factors such as extroversion and sensitivity, steady-state traits such as anxiety, transient states such as motivation and feelings of self-efficacy, as well as language learning aptitude, as influences underlying (causing) differences in language learning. We will examine a sub-set of these factors in turn.

Aptitude

The notion of language learning aptitude is one that applies most aptly to adult second language learning, especially in formal instruction contexts. The construct of aptitude is, however, a slippery one, for it involves the postulation of underlying cognitive and motor sub-skills that are independent of prior language learning experience. In relating aptitude to language gain, it is essential to recognise that the cognitive and motor sub-skills must logically exist prior to the onset of individual differences in outcomes. Aptitude, in other words, is a predictive construct.

Aptitude is operationalised — that is, defined through an overt measurement tool — through specific components which in combination create ‘aptitude’. The most successful and widely used measurement tool in language learning is the Modern Language Aptitude Test — MLAT (Carroll and Sapon 1959), which is comprised of five sub-tests. These are:

- the Number Learning Task designed to assess ‘auditory alertness’;
- Phonetic Script Identification Task, which requires sound-symbol matching;
- Spelling Cue Task in which vowels are deleted from key words. The candidates must then associate the deformed key word with an English synonym;
- the Words in Sentences sub-test of the MLAT, in which syntactic functions of sample sentences are underlined. Different sentences with similar word types in differing functional roles are underlined in the distractor set. Here, the task is to match the syntactic role in the key to the same role in the alternatives;
• the Paired Associates task which requires a time-constrained rote memorisation of Kurdish words paired with English cognates.

The research literature on the MLAT suggests that it is positively and often highly correlated with foreign language learning success (see Ehrman 1996 for a review). Its usefulness, however, has been limited by the fact that it requires knowledge of English as a precondition of administration, making it generally inappropriate for non-native speakers of English. The construct of language learning aptitude, if it is psychologically universal, should be valid across the particulars of any reliable measurement instrument constructed to instantiate it. Sasaki (1996), for instance, has created an aptitude test for native speakers of Japanese that appears to manifest a degree of validity comparable to the original MLAT battery.

The influence of aptitude in uninstructed adult second language acquisition is not well known. One view, held by Bley-Vroman (1989), is that differences in adult SLA outcomes are attributable to individual differences in such cognitive skills and strategies as problem-solving ability, distributional analysis and hypothesis formation. The essential problem here is the operationalisation of these constructs into reliable and valid measurements — since none appear to be available, Bley-Vroman’s hypothesis has to date gone untested. At the time of the present review, the most reliable measurement tools used for accessing some of the constructs Bley-Vroman alludes to are those already mentioned which define aptitude as a configuration of linguistic sensitivity factors — those that tap into individuals’ ability to make sense of decontextualised language (Skehan 1989; 1991) as instantiated in the MLAT.

In one of the largest studies done on SLA in a naturalistic context, Ginsberg (1992) examined a large database compiled by the American Council of Teachers of Russian (ACTR) over a period of nearly 20 years. Three research issues related to language gain were examined: (1) interrelationships among pre-program language measures (standardised listening and reading proficiency tests, the OPI, and the grammar-based ACTR qualifying exams); (2) predictors of gain on various criteria, with particular focus on gender, knowledge of other foreign languages, other individual attributes and characteristics of previous language learning careers, the MLAT, grammar-based ACTR qualifying exams, and proficiency in other language modalities; and (3) interrelationships among gains in different modalities.

Brecht (1993) reports on different aspects of the same data. Both Ginsberg (1992) and Brecht review the results of this large-scale study of 658 American college and graduate students studying Russian in the former Soviet Union between spring 1984 and spring 1990. Variables examined included student characteristics (age, gender, citizenship, country of birth, place and levels of formal education, highest degree taken, major, prior Russian and other second language training and experience, program type, overseas host institution, and a variety of language measures (proficiency tests, program qualifying exams, learning style or aptitude data). Brecht found that certain student characteristics were predictive of language gains abroad, including gender, experience in learning other foreign languages, and command of foreign language grammar and reading skills prior to
exposure to Russian in a naturalistic context. The researchers also found that the MLAT was a valid predictor of success in developing reading and listening, but did not predict oral proficiency gains well. Brecht, Davidson and Ginsberg (1995) note that this phenomenon may reflect the commonality between the cognitive tasks presented on the MLAT sub-tests, and the kinds of tasks that are characteristic of formal language instruction (Larsen-Freeman and Long 1991). The predictive validity of MLAT appears to diminish when learners are exposed to unstructured language acquisition contexts. These findings, however, are not entirely clear. The best predictor of listening gain was MLAT 3 (the Spelling Cue Task in which vowels are deleted from common English words, and then associated with an English synonym). It is puzzling that this sub-skill would correlate with listening gain in Russian, but not to oral proficiency gain, especially when we consider the interrelatedness of listening and speaking abilities.

Age

Maturational constraints have long been known as fundamentally important factors in adult second language acquisition success (Singleton 1989; Long 1990b). Palmer’s (1989) study of Yemeni adults, for instance, found an impact for age — that, as expected, there is a clear negative correlation between age and TOEFL gain. He notes also that older learners are more likely to be parents — with less time available for full-time study than their younger peers might have, making his results potentially confounded. Still, in terms of cross-sectional research with large groups of adults, the pattern appears to be quite stable, and is in fact one of the empirical pillars of what Long (1990a) concludes is a ‘set of laws’ which require explanation by a theory of SLA (cf Bialystok 1997).

Numerous explanations for one of the most commonly observable phenomena in adult SLA have been formulated, ranging from social psychological constraints (Schumann 1978) to physiological constraints such as changes in neural plasticity (Scovel 1988; Birdsong 1992). Here again, the generalisation that ‘younger is better’ in second language acquisition in naturalistic exposure contexts appears to hold true. In formal learning contexts, however, there appears to be less of an advantage for children, owing to the better overall cognitive development of adults.

For both children and adult second language learners, there is a potentially confounding variable which often interacts with age — that is, the start and duration of exposure to the second language. Younger learners may not in fact have an advantage if their exposure is sporadic or interrupted. This leads us to another important factor in understanding second language gain — the extent and type of exposure.

Exposure to the second language

Stern (1976), in reviewing Canadian immersion programs, contested the finding that ‘pupils taught French from the age of eight do not
subsequently reveal any “substantial” gain in achievement when compared to students who start three years later. This phenomenon is explained in terms of environment and time, not age, and contrasts with the thesis that younger language learners are more efficient. With respect to an optimal age for learning a second language, it seemed clear to Stern that early learning as such, especially in small doses, does not guarantee success.

Walberg (1978) conducted research on Japanese children in the USA to test the hypothesis of early age sensitivity in second language learning. The results did not support this hypothesis. Acquisition proceeded at a fast rate initially, but the amounts of gain diminished with time. Walberg’s finding suggests that children’s eventual settlement into ethnic enclaves, which may feature first language maintenance instruction and home language use, may serve to create a picture of relatively slower rates of second language acquisition, even for children.

The situation for adults in a foreign language context is quite different. While exposure in terms of years of residence may be mediated by the extent of enclosure in an ethnic enclave for adults in an unstructured SLA context, for formal language learning — instructed SLA (Ellis 1990) — hours of classroom instruction tend to provide the best predictor of language gain.

This difference is highlighted in a study of language gain among Spanish speaking migrants to the USA. Mooney (1991) hypothesised that a regression model made up of the variables of instrumental motivation, age, educational attainment (years of previous schooling), exposure to English (length of residence in the USA) and informal use of English would account for significant amounts of variation among ESL test scores.

The New York State Placement Test for English as a Second Language Adult Students was used to assess oral proficiency of 100 Limited English Proficient (LEP) adults at 0 hours, 50 hours, and 100 hours of instruction. Data which was collected during an intake interview consisted of information regarding student demographics and educational background. Information regarding instrumental motivation (motivation characterised by students’ desire to learn a second language for instrumental reasons such as to obtain or upgrade employment) and informal use of English was also obtained. The learner variables were examined to detect relationships between them and language performance (Test 1, Test 2 and Test 3). These factors were also studied regarding their capacity to predict achievement at three intervals: (1) Gain 1 (between 0 and 50 hours) (2) Gain 2 (between 50 and 100 hours) and (3) Gain 3 (between 0 and 100 hours). Four additional variables (gender, reading and writing ability in Spanish and employment status) were also examined to detect relationships to language performance and achievement.

Through a hierarchical multiple regression analysis, Mooney found that age, educational attainment (years of previous schooling), and informal use of English and gender were significant predictors of achievement. When all five variables were considered together in a single model, educational attainment was the only significant predictor for gain occurring between
Mooney’s study reveals that gain is most rapid for adults who have had most education in their home country, although informal use (exposure) is a significant contributor to individual differences. If we consider exposure as opportunity to learn, we can examine the impact of time in formal language learning contexts. Two studies done in Japan are of interest in this regard.

Saegusa (1985) examined gain scores for 1173 pre- and post-instructed Japanese company employees taking the Test of English for International Communication (TOEIC). Not surprisingly, the amount of classroom instructional time contributed to significant gain differences among the candidates for both the skills assessed on the TOEIC — listening and reading. Saegusa concluded that it would take in excess of 400 hours of classroom instruction to effect a change of 150 standardised scale score points — roughly equivalent to a change from an OPI rating of ‘1’ (survival level) to an OPI rating of ‘2’ (minimum working proficiency).

In a more recent study of gain on TOEIC, Boldt and Ross (1998) gathered archival files from 23 Japanese corporations with in-house English language training programs. They sampled 1147 pre- and post-instruction test score records with a view to encoding characteristics of the language programs and hours of exposure to classroom instruction in both neural networks and linear regression models of language gain. Boldt and Ross found that in addition to the sheer amount of language tuition in hours, some curricular aspects of the language programs themselves (such as teacher qualification and in-service training, and use of authentic materials) contributed to differential impact for language learning gain on this standardised test. A further issue examined in their study was that of instructional intensity. In addition to the sheer amount of instruction, the intensity, as measured by hours per week in intensive or extensive instruction, was examined. Figure 6.1 shows the relation of exposure and intensity of exposure.

*Figure 6.1: TOEIC gain as a function of exposure and intensity*
As would be expected, gain on this standardised test is mainly a function of exposure (hours of instruction). There is also a beneficial effect for extensive exposure relative to intensive instruction over short periods of time.

Exposure can been seen as a key variable in competency-based assessment schemes as well. Ross (1997) analysed certificate achievements in the Certificate of Spoken and Written English (CSWE) (Burrows et al 1996), the curriculum framework used within the national Adult Migrant English Program (AMEP) in Australia. He found that exposure in the form of hours of classroom tuition was the most stable predictor of CSWE competency gains. As hours of instruction increase, the probability that the adult migrant candidates will achieve a CSWE award should increase monotonically. This is what in fact occurs.

**Figure 6.2: CSWE gain probabilities**

Ross noted that in CSWE Stage 1, banding categories (slow, average and fast-track learner profiles) maintained their expected ordering, but also found that the expected category orderings were inconsistent in Stages 2 and 3. This finding suggests that individual differences that motivate banding in Stage 1 are not adequate predictors of learning pace in later stages. An example of the probabilistic model Ross derived for CSWE Stage 2 is given in Figure 6.2 above.
Chapter 7

Language gain
measurement criteria

The notion of language gain can be separated into different inferential tasks for the researcher or program evaluator. The first is in the marshalling of evidence that there has indeed been some form of non-trivial language gain. Second is that the observed gains are not artefactual or ephemeral and that these observed gains are therefore reliable/dependable and valid. The third and most difficult inferential task involves attributing causative status to discrete factors – such as the language instruction process. The remainder of this section outlines methods supporting these three types of inference about language gain.

Evidence of gain

The first inferential task is to determine that there has been change (gain). As stated earlier in this monograph, fundamental to all gain studies is a pre-test and a post-test. We assume that the two measures are parallel forms such that the mean of the pre-intervention test (X) is equal to the mean of the post-intervention test (Y). We also must assume that the respective variances (here, standard deviations) are also equal; Sx = Sy. Our first line of evidence that there has been gain is in the observation, through repeated measures of the same persons, that the means of the pre- and post-tests are not equivalent, and that y>x. We of course assume this after intervention (instruction). Further, we can examine the average change with a paired (repeated measures) t-test (Guilford and Fruchter 1978) in order to estimate the probability that the observed difference could have emerged by random chance alone.

Figure 7.1: Paired t-test

\[
t = \frac{\sum D}{\sqrt{\frac{N(\sum D^2) - (\sum D)^2}{N - 1}}}
\]

where \(D\) = the observed difference of \(Y-X\) for each person.

A significant paired t would indicate that there has been change or gain in the score, provided that \(Y\) is greater than \(X\). Such a comparison satisfies the first criterion for measurement of gain — that there has been a gain
that is unlikely to have occurred by chance alone. The reason for the gain, in lieu of a control group, must remain only speculative.

Two other important points are worth considering in this sort of gain score analysis. The first is the impact of subject attrition. It may be misleading if the post-intervention test scores are available only for the ‘survivors’ of the program. If there is systematic attrition (drop-outs) by the least successful participants in the program, the inference that there has been ‘gain’ must be constrained by the possibility that only the most motivated and successful candidates presented themselves for post-testing. The ‘gain’ is thus limited to the sub-set of participants who were perhaps likely at the outset to benefit from the program. This selective post-testing threat needs to be considered in the interpretation of such non-causeative designs for establishing gain.

The second point to consider is the assumption we make that the pre-test and post-test are not only parallel, but are equally internally consistent measures of the same construct. Without evidence of their internal consistency (discussed below as ‘reliability’), the inference that there has been gain may be misleading.

**Test score reliability and dependability**

Since observations of language performances across time are fallible, they are subject to variation. This variation is fundamental to all assessment, and is the notion underlying the terms ‘generalisability’, ‘dependability’ or ‘reliability’. For language gain, if either of the two observations in time are subject to instability, the inference that there has been gain can be a mistaken one. The methods of assessing the reliability of performances match the types of measurement systems used. The reliability of each of the three main types of language gain approaches will be reviewed here.

When standardised (norm-referenced or NRT) scores are used, the individual assessments are most often atomised components of language knowledge or ‘items’ that are hypothesised to form the basis of the trait of interest (for example, second language reading). The notion of reliability here is that an observation at one time will remain stable unless there is some intervention (learning) that serves to render the state of an individual’s knowledge unstable. The fundamental definition of reliability here is that a test followed by an immediate retest would result in an identical rank order of individuals on that trait (see Henning 1987 or Brown 1996 for fuller discussion). This notion is readily expressed in a direct correlation between the test and its immediate replication. Of course, logistical constraints make this design very rarely used in practice. In its stead, an indirect method of estimating the average correlation among the parts of the test (the items) is commonly used. It is by inference that when we observe high internal consistency among the items on a language construct, there is stability in the rank order of observed performances. Empirical tests of internal consistency with direct test-retest methods support this inference (Swinton 1983).
In the usual context of a gain score study, immediately parallel pre-testing is not feasible. In this case, internal consistency can be estimated by examination of the item variances (or by split-half correlational approaches). The assumption of internal consistency holds for both the pre-intervention and post-intervention measures, for if one or both of them are not internally consistent, the evidence that there has been gain crumbles, because the construct itself (for example, gain in reading ability) may not have been adequately measured. Figure 7.2 gives the Kuder-Richardson approach to internal consistency estimation.

**Figure 7.2: Internal consistency**

\[
\frac{n}{n-1} \left( \frac{S^2_t - \Sigma S^2_i}{S^2_i} \right) = \Gamma_{xx}
\]

where

- \( n \) = the number of items
- \( S^2_t \) = the variance in total scores
- \( S^2_i \) = the variance in items

Stability of scores at one time (reliability) is the basis for hypothesising that a change in a scaled score at a later time indicates ‘true’ score change for the individual. Again, since the scores are all relative to each other, a higher scaled score later in time (after instruction or natural acquisition) implies that the individual has gained on the trait of interest relative to the entire population of language learners. Although the items are only indirect indicators of the underlying component microskills thought to constitute the second language system, changes in an individual’s scaled score imply gain relative to the population as well as absolute gain in language knowledge.

In many assessments, indirect measures of abstract constructs might not be used. Rather, direct performances (McNamara 1996), because they have greater face validity and sample more authentically the requisite language performance tasks, are more commonly being used for assessment. Here we have a different requirement for establishing reliability. Since the performances must be refereed by competent expert judges, we have a new source of variance — differences among the judges and, potentially, differences among the possible tasks assigned to candidates. In performance assessment we are concerned that the individual differences we observe indicate true differences between the candidates, and are not because of differences in referee severity or task complexity.

Generalisability Theory (Brennan 1983; see also Lumley, Lynch and McNamara 1994) can inform assessment designers about the sources and severity of dependability shortcomings for competency assessments. A G-study and associated decision study (D-study), are essentially one-off experiments whereby a representative cohort of candidates is assessed with a series of representative tasks, which are then rated by a team of representative judges. The aim is to identify which facets are the source of variance in the assessment process. The D-study informs designers how many tasks and judges are needed.
Here the notion of stability of an observation is referred to as its ‘generalisability’ — how adequately the sampled tasks and judges generalise to the world outside the assessment venue. The approach used for performance assessment requires an estimate of how much tasks vary in their difficulty (analogous to internal consistency), and also, the degree to which different judges can be expected to vary in their assessments of those performances. Figure 7.3 introduces the Generalisability Coefficient (Brennan 1983), which estimates the variance components associated with the task and judge facets in a performance assessment.

**Figure 7.3: Generalisability coefficient**

\[
\frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_t^2}{n_t} + \frac{\sigma_j^2}{n_j} + \frac{\sigma_t^2}{n_t} \cdot \frac{\sigma_j^2}{n_j}} = G
\]

where \( \sigma_p^2 \) is the person variance
\( \sigma_t^2 \) is the task variance component
\( \sigma_j^2 \) is the judge variance component

In the event that tasks are equivalent and judges do not vary in their consistency of rating person performances on those tasks, the only major variance component is that by which the inference of language proficiency is made — the person facet (or more correctly, how individuals differ in their ability to perform the task).

**Gain score reliability and dependability**

Changes in language performances over time imply that there has been gain. Yet this inference is dependent on the reliability or generalisability of the tests or performance observations at both times in the language gain measurement process. In order to estimate the reliability of the gain score, provided that there is a significant gain (Figure 7.1), it is necessary to know the internal consistency for both the pre-intervention and post-intervention measurements. These can be used in combination with the correlation between the pre- and post-intervention scores (Suen 1990). Figure 7.4 shows this difference (or gain) score reliability.

**Figure 7.4: Gain score reliability**

\[
r_{dd} = \frac{r_{xx} + r_{yy} \cdot 2r_{xy}}{2 - 2r_{xy}}
\]

where \( r_{xx} \) is the internal consistency of the pre-test
\( r_{yy} \) is the internal consistency of the post-test
\( r_{xy} \) is the product moment correlation between the two tests
Zimmerman and Williams (1982), Rogosa and Willet (1983), and Williams, Zimmerman and Mazzagatti (1987) discuss modifications of the internal consistency and correlation-based approach to gain reliability that are optimally sensitive to changes in score distributions from pre-instruction to post-instruction. They add ratios of standard deviation terms (theta, in Figure 7.5 below) to make a product of internal consistency and the ratio of pre-test and post-test score distributions. The Zimmerman and Williams modification (Figure 7.5) makes explicit the assumption that greater variation among learners is expected before instruction relative to variation after instruction. Figure 7.5 shows the Zimmerman and Williams (1982) modification of the internal consistency and correlation-based gain reliability.

**Figure 7.5: Gain score reliability with pre-post variances**

\[
\text{Gain reliability} = \frac{\lambda_1 r_{xx} + \lambda_2 r_{yy} - 2 r_{xy}}{\lambda_1 + \lambda_2 - 2 r_{xy}}
\]

where lambda 1 is the ratio of pre-to-post-test standard deviations
lambda 2 is the ratio of post-to-pre-test standard deviations
\( r_{xx} \) is the internal consistency of the pre-test
\( r_{yy} \) is the internal consistency of the post-test
\( r_{xy} \) is the product-moment correlation between the two tests

An assumption of this and other gain score reliability indices is that there will be variation in the impact of the program. Namely, there will be differential learning among the program participants. The reliability of gain scores will be highest when there is a 'counterfactual' or control group design (Mohr 1992). Non-recipients would be expected not to gain at the same rate as the recipients of the program intervention, leading to a change in the ordering of the participants relative to the non-participants. As the program takes effect, recipients may gain differentially within their own group, and as a group, should gain at a much faster rate than the non-recipients. In the event that a counterfactual group is not included in the gain score study design, the caveat noted above — that attrition should be avoided, and for there to be incentives to bring in even marginal participants for the post-testing — becomes even more crucial for reliable gain score interpretation.

In order for the gain score reliability concept to apply to criterion-referenced tests, where decisions are absolute (pass or fail), some adaptations to the above approaches are needed. By replacing the internal consistency estimates for the pre-instruction administration of the criterion-referenced test with a squared-error loss agreement coefficient phi (see Brennan 1983; 1984 and Brown 1996), fixed at a cut score lambda for each of the pre- and post-test administrations, the Williams, Zimmerman and Mazzagatti (1987) approach can be adapted to assess gain score dependability. Here, pre-instruction criterion-referenced measures are used as a baseline for language learning gains as indicated on post-instructional criterion-referenced measures for the same cohort of
learners. This approach is premised on there being a cut score on both the pre-instruction and post-instruction versions of the criterion-referenced tests. Figure 7.6 shows the modification of the norm-referenced approach to gain score reliability to suit the conditions of criterion-referenced gain score dependability.

Figure 7.6: Gain dependability with cutting scores

\[
\text{Gain dependability} = \frac{(\lambda_1 f(\theta_1)) + (\lambda_2 f(\theta_2)) - 2r_{xy}}{\lambda_1 + \lambda_2 - 2r_{xy}}
\]

where \(\theta_1\) is the ratio of pre-to-post-test standard deviations
\(\theta_2\) is the ratio of post-to-pre-test standard deviations
\(\phi\) (\(\lambda_1\)) is the squared-error loss agreement on the pre-test
\(\phi\) (\(\lambda_2\)) is the squared-error loss agreement on the post-test
\(r_{xy}\) is the product-moment correlation between the two tests

Ross and Hua (1994) used this approach to determine the optimal index of gain score dependability in a pre-instruction and post-instruction study to assess the language learning gains. They examined the cut score dependability of the pre-instructional administration of a criterion-referenced English for academic purposes listening test as well as a parallel form in a post-instructional criterion-referenced test. Ross and Hua also examined the differences in the ratio of pre- and post-instruction variances. Their results indicated only moderate gain score dependability, because the cut score was close to the mean proportion correct on the post-instructional test. They concluded that a counterfactual group design would provide a higher dependability index, and that a period of program intervention longer than the single semester sampled would most likely lead to higher gain dependability.

Gain score validity

Rarely does a demonstration of gain reliability or dependability provide a sufficient demonstration of a program’s effectiveness. The question most often asked is: ‘What does the gain mean outside the context of the instruction or intervention?’ Here, the need for demonstrating the validity of gain scores becomes crucial.

In order to claim that learners have acquired (gained) a specific trait (for example, self-confidence), there would have to be two observations of the trait. However, just because we have called one a survey instrument (for example, ‘Self-confidence Profile’), it does not follow that the instrument in fact provides valid information about learner self-confidence. The instruments created to assess gain on specific traits need to be validated. Apparent gain may be an artefact of the measurement instruments. When repeated measures of the same task are used, the gain may be of familiarity with doing that kind of task, not of stable proficiency gain. For this reason, we need to provide evidence that individual differences in gain go beyond
the particular tasks we have chosen, and extend to a same-construct external criterion.

Since the notion of language gain includes a temporal dimension whereby proficiency measured at one point in time changes according to subsequent individual differences in learning, motivation, attention, attendance and so on, the expectation for valid assessment of gain is that individual differences observed early in the instructional period should be less correlated with assessments of achievement and proficiency administered late in the instructional period. Late-administered assessments should be more correlated among themselves than with the earlier assessments. If there is little intersubject variation in gain, that is, that all learners gain at equal rates, their relative proficiency rankings can be expected to remain the same. In such an instance, the pre-post test correlation is expected to be very high, as would the observed t (Figure 7). The usual case is, however, that there will be considerable variation in the impact of the program because of individual differences in participation. With the foregoing caveat that post-testing should avoid attrition, we can examine the validity of gain with the use of parallel post-tests. Gain validity is supported by the observation of correlations of differential magnitudes. The correlation between pre-test and post-test (parallel forms: \( r_{xy} \)) should be of moderate magnitude if there is some change in the rank ordering of program participants (especially if a control group is included in the design). The pre-post correlation magnitude should be approximately equal to the pre-external test correlation: \( r_{xz} \). If the gains are valid, we would expect to see the largest correlation between the post-test and external criterion, which co-vary in terms of time (both are post-intervention), and construct (both are measures of the same thing): \( r_{yz} \). We thus hypothesise that for validity of gain over time, we would observe \( r_{xy} = r_{xz} < r_{yz} \).

An example of this approach can be seen in the matrix below. Here, TOEFL listening sub-tests are compared in the pre-post-test manner described above. At the end of the instructional period, all learners are given an external (Comprehensive English Language Test, Listening sub-test: CELTAL) in addition to the TOEFL parallel form of the post-test (LC2). Evidence of gain validity is gleaned from the different magnitudes of the observed correlations.

<table>
<thead>
<tr>
<th></th>
<th>TOEFL pre-test</th>
<th>Instruction</th>
<th>TOEFL post-test + CELT listening</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LC1</strong></td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LC2</strong></td>
<td>0.671</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td><strong>CELTAL</strong></td>
<td>0.670</td>
<td>0.753</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Another example serves to explicate the different uses of this approach. Here, syllabus-sensitive achievement assessments (AT), and cumulative teacher holistic evaluations of learner performances (HE) are used as intermediate assessment tools. A non-syllabus related proficiency test...
(TOEIC) is used in the usual pre-post-test configuration.

<table>
<thead>
<tr>
<th></th>
<th>HE</th>
<th>AT</th>
<th>T1</th>
<th>T2</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td>0.411</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>0.277</td>
<td>0.442</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>0.496</td>
<td>0.630</td>
<td>0.644</td>
<td>1.000</td>
</tr>
</tbody>
</table>

The correlations of holistic evaluation (HE) and achievement (AT) with TOEIC post-test (T2) are larger than the correlations of HE and AT with TOEIC pre-test (T1); that is, 0.496 > 0.277 for HE; 0.630 > 0.442 for AT. This implies that learners eventually differ in their gain in proficiency, and that the difference is consistently assessed because HE, AT and T2 occur closer in time.

In a more sophisticated model than the simple correlational approach shown above, Gupta, Srivastava and Sharma (1988; 1989) define gain score validity as one based on relative magnitudes of pre-post-external criterion correlations, where the external criterion is a same-construct (parallel form, or ideally, different method of measurement, same-construct auxiliary measure). Figure 7.7 shows their approach to gain score validity estimation.

**Figure 7.7: Gain score validity**

\[
\text{Gain Val} = \frac{r_{xz} - (r_{xz} \lambda_1)}{\sqrt{(\lambda_1^2 - (r_{xy} \lambda_1)) + 1}}
\]

Where

\[
\lambda_1 = \frac{S_{\text{pre}}}{S_{\text{post}}} \quad \lambda_2 = \frac{S_{\text{post}}}{S_{\text{pre}}} \quad \text{standard deviation ratios}
\]

The Gupta et al approach offers the advantage of including ratios of standard deviations, which should be sensitive to the variation (spread) of scores on the pre-to-post comparisons. Homogenous gain would lead to smaller spread on the post-test than on the pre-test.

**Gain score causation**

Even if we have evidence of reliable and valid gain, we are at a loss to infer why there has been gain. Ideally, we would use a true experimental design, in which randomisation to the intervention or counterfactual (control) group would ensure a distribution of random error. Rarely can we do this, however. As an alternative, we can often use a quasi-experimental design.
with non-equivalent groups analysis of co-variance (ANCOVA). Figure 7.8 sketches the design of a quasi-experimental design using ANCOVA.

**Figure 7.8: Venn diagrams of ANCOVA design**

Post-test(y)  Pre-test (a)  Group  (b)

1 - r²  

R²_{ab} - R²_a

We would test the significance of the grouping variable (membership in program or control groups) on the post-test variance that exists after the pre-test variance has been partialled out. Recall that with a control group design, we expect the pre-post correlation to be smaller as time and the impact of the intervention create differences among members of the groups.

The object of interest in establishing plausible causes of the differences observed is the post-test variance. Here we would hypothesise that the grouping variable will have a large correlation with the post-test and a near zero correlation with the pre-test. When we observe this pattern, we may infer that the group-post-test correlation is large because of the impact of the program or intervention. This hypothesis is tested with the F-distribution (Huitema 1980; Cohen and Cohen 1983). If the observed F (Figure 7.9) is larger than the tabled F, given the degrees of freedom, we may infer that the program impact is unlikely to have come about by chance — specifically because the group differences for program versus no program emerge after intervention, and because the moderating impact of the pre-test (or other co-variates) have been controlled in the model.

**Figure 7.9: Fisher’s test for ANCOVA**

\[
F = \frac{R_{YAB}^2 - R_{YA}^2/KB}{1 - \frac{R_{YAB}^2}{(N \cdot KA \cdot KB \cdot 1)}}
\]

where \(R_{YAB}\) is the multiple correlation of pre-test (A) and group (B) with the post-test (y)

\(R_{YA}\) is the correlation between pre-test (A) and post-test (y)

\(K_A\) is the number of pre-tests (co-variates) in the model

\(K_B\) is the number of groups in the design

\(N\) is the total number of persons in all groups
Differential group gains

It is not unusual for researchers in language gain to be interested in group differences. Given the types of factors affecting differential language learning rates (outlined in Chapter 1), it is often important to examine how much gain is affected. Often, program policy is predicated on the concept of learner symmetry — which assumes that if all program participants are given the same amount of tuition, they should in principle achieve the same rate of gain.

Research into individual differences should serve to temper this view with the expectation that there will be differences in outcomes. The concept of group differences, however, requires a group-based analytic strategy in order to investigate differences in language gain.

In order to examine putative group differences, there needs to be a coherent operationalisation of group that subsumes the concept of individual differences. That is, natural ‘groups’ need to include the same variation attributable to individual differences within the larger groups identified. An example of putative natural grouping would be a factor such as language typology. Learners who already know an Indo-European native language would constitute one group, while a different group might be persons who have acquired a Sino-Tibetan first language. Within each ‘group’ we would expect a similar range of individual differences in language learning aptitude and motivation.

Analogous to the analysis of co-variance approach introduced above, the standardised change score analysis (SCSA) approach (Huitema 1980; Lynch 1996) begins with a pre-test given before program intervention. Here, however, participants’ membership in the naturally formed group is used to create a new variable, which contains a dichotomous code for membership. In contrast, in the ANCOVA, ‘groups’ indicate participation/non-participation in the program.

In the SSCA design, the grouping variable is correlated with the pre-test. In the ‘ideal’ impact study, one in which all participants start out equally, the correlation between the group code and the pre-test ($r_{xd}$) should be very small. Likewise, if the impact of the intervention leads to symmetric gain, we would expect to observe two phenomena: (1) that there would be a large observed paired $t$ (Figure 7.1) when individuals are compared pre- to post-intervention; (2) the correlation between the group code and the post-test ($r_{yd}$) will be of the same magnitude as the pre-test group correlation. This outcome would imply that changes in scores from pre- to post-test are group independent — that is, they are symmetric.

The usual application of the SCSA design is motivated on the hypothesis that there is differential group gain. For this reason, the group codings chosen need to be indicative of groupings that are clearly hypothesised to be factors underlying differential outcomes in language instruction programs. In order to test the hypothesis that there is differential gain, we can again resort to the two-step process given above: (1) that the paired $t$-test (Figure 7.1) will indicate overall gain within individuals, but that (2)
there will also be a large Hotelling-Williams t (Figure 7.10), which is a formal test that the pre-post group correlations are equal.

**Figure 7.10: Hotelling-Williams t-test**

\[
t = (r_{xd} - r_{yd}) \cdot \sqrt{\frac{N - 1}{N} \cdot \frac{(1 + r_{xy})}{2}} \left\{ \frac{N - 1}{N - 3} \cdot R + \left[ r^2 \cdot (1 - r_{xy})^3 \right] \right\}
\]

where \( R = \left[ 1 - r_{xy}^2 - r_{xd}^2 - r_{yd}^2 \right] + \left( 2 \cdot r_{xy} \cdot r_{yd} \cdot r_{xd} \right) \)

\( r^2 \) is \( (r_{xd} + r_{yd})/2 \)

and \( (r_{xy}) \) is the pre-post-test correlation

In the event we find that condition (1) above holds, and that the Hotelling-Williams t-test is not significant, we can infer that there is no group-dependent asymmetry in the program’s impact.

The usual situation in instructional programs, however, is that members in different natural groups ‘fan out’ in their response to instruction in such a manner that there is considerable variance near the end of the instructional period. In such an outcome, the Hotelling-Williams t-test of group differences in the SCSA design would be highly significant. The implied ‘fan spread’ phenomenon indicates, especially in adult migrant programs, that different groups of adult language learners gain in language (and possibly non-language) outcomes at different rates. It often suggests a selection by maturation interaction which produces a ‘fan spread’ whereby groups grow apart while individual differences within groups increase (Kenny 1979). We thus have two simultaneous factors mediating the impact of the program — individual differences within groups and between group factors. Presented graphically, this phenomenon looks something like that in Figure 7.11.

**Figure 7.11: Fan spread in language gain**
Here, the group effects (dark lines) widen as the program intervention increases. Concurrently, individual differences within groups (narrow lines) also ‘fan out’ near the end of the program.

Policy related to an anticipated rate of gain (as assessed by a tally of achieved competencies, for example) may need to accommodate the fan spread phenomenon by projecting group-specific level changes according to a profile method. If policy makers anticipate that all groups learn at the same rate, some groups may literally ‘run out’ of tuition before they reach the desired outcome level. Other groups may reach the desired threshold ahead of schedule.

**Multi-wave gain analysis**

Since symmetric gain in both language and non-language outcomes cannot be realistically expected across all adult candidates, simple ‘two-wave’ before-and-after change models in many cases fail to provide an adequate picture of the gain process. Nor do they provide an adequate intervention strategy that optimises symmetric long-term outcomes. Multi-wave assessments of gain, based on dependable and valid instrumentation, are what is needed (Rogosa 1995). This type of assessment procedure could integrate diagnostic features into the process. For instance, group-dependent differences could be diagnosed early in the program intervention, thereby leading to an adjustment of the tuition scheduling for the groups with the flattest learning curve.

A graphic representation of individual differences (Figure 7.12) based on a multi-wave approach reveals how adult language learners show differential rates of learning which are hidden in aggregate, group-based analyses.

*Figure 7.12: Multi-wave observation of SLA gain*
Individuals may backslide, regress to the mean, manifest ‘U-shaped’ learning curves, and hit intermediate plateaus before eventual interlanguage revision and subsequent gain. Although obviously more expensive and time-consuming than the pre-post design, the multi-wave approach to examining language gain shows the potential for providing the clearest explanatory picture.

**Time-lagged gain analysis**

The final approach to gain analysis introduced here is an observation of gain in a time-lagged repeated measures design. Here, cross-sectional samples of language learners of different backgrounds are assessed at two or more times in their language learning careers. Two (or more) measures (rectangles) are used to indicate proficiency at a given stage. Parallel measures are used at later stages as repeated measures. The latent outcomes (ovals) may vary in relation to each other over time if changes in proficiency are asymmetric among the same group of learners. Changes in the magnitude of standardised path coefficient from one time to the next are the object of interest here (Loehlin 1987). Contrastive groups of learners can be assessed using the same comparative index in order to examine differential patterns of influences on change from one time to another. Figure 7.13 represents the measurement model for a time-lagged, or ‘panel’ study of changes over time.

**Figure 7.13: Time-lagged panel design**

In such a longitudinal study, contrastive groups would be assessed with the same instruments or tasks (denoted with rectangles) as the ‘non-Indo-European’ language group sample. The object of interest would be whether there are different magnitudes of paths (Cohen and Cohen 1983) from one time (T1–3) to another as indicated by the time-lagged measures. Different path magnitudes across the contrastive samples would suggest variances in the impact of time (and instruction) on the process of SLA in the
contrastive groups. Thus, the gain researcher could construe that ‘gain’
does not develop symmetrically across the contrasted groups, and that
other mediating factors are likely to exist to suppress the direct influences
of instruction and naturalistic second language acquisition.
Conclusions

This monograph has reviewed a number of issues in the assessment of language gain. These revolve crucially around the concepts related to the coherent measurement of language abilities — both relatively (in the norm-referenced sense) and absolutely (in the competency or criterion-referenced sense). In either approach, there is a common requirement for inferences about language gain to stand up to critical analysis. The first necessary condition for this requirement to be met is that observations of ability at the earliest point in time are of necessity highly consistent, stable, dependable, and generalisable — that is, that the first observation must be reliable. The second condition subsumes the first. The assessment of ‘outcomes’, be they naturalistic or the product of instructional intervention, need to be both reliable and valid. This is to say that observed outcomes need to be indicative of abilities, competencies, knowledge, skills and the like, which are operational in the society outside the immediate context of the intervention. These two requirements will constitute a significant challenge for program evaluators hoping to muster convincing evidence of language gain.
Bibliography


Brindley, G and S Ross. 1998. 'A multiple method approach to language test validation'. Unpublished manuscript.


Freed, B 1995. What makes us think that students who study abroad become fluent? In B Freed (ed). Second language acquisition in a study abroad context. Amsterdam: John Benjamins

Gardner, R 1985. The social psychology of second language learning: The role of attitude and motivation. London: Edward Arnold


German, P and K Perkins 1984. ‘Cloze residual gain: A procedure for measuring information gain in ESL reading comprehension’. In Language Key to Learning. Selected Papers from the Annual State Convention of the Illinois Teachers of English to Speakers of Other Languages/Bilingual Education


Huitema, B 1980. The analysis of covariance and alternatives. New York: John Wiley and Sons


Rogosa, K L and J B Willet 1983. ‘Demonstrating the reliability of the difference score in the measurement of change’. Journal of Educational Measurement, 20, 4: 335–343


Spolsky, B 1990. 'The prehistory of TOEFL'. *Language Testing, 7, 1*: 98–118


Walberg, H 1978. 'English acquisition as a diminishing function of experience rather than age'. *TESOL Quarterly, 12, 4*: 427–37

Weissberg, R C and M Stuve 1979. 'Differential gain rates in intensive ESL programs — Who gains the most?' *System, 7, 1*: 61–65


Zimmerman, D W and R H Williams 1982. 'Gain scores in research can be highly reliable'. *Journal of Educational Measurement, 9, 2*: 149–154

Bibliography 55
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTR</td>
<td>American Council of Teachers of Russian</td>
</tr>
<tr>
<td>AMEP</td>
<td>Adult Migrant English Program</td>
</tr>
<tr>
<td>AMES</td>
<td>Adult Migrant English Service</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>Non-equivalent Groups Analysis of Co-variance</td>
</tr>
<tr>
<td>AT</td>
<td>Achievement</td>
</tr>
<tr>
<td>CAT</td>
<td>Computer Adaptive Test</td>
</tr>
<tr>
<td>CELT</td>
<td>Comprehensive English Language Test</td>
</tr>
<tr>
<td>CELTAL</td>
<td>Comprehensive English Language Test, Listening sub-test</td>
</tr>
<tr>
<td>CSWE</td>
<td>Certificate of Spoken and Written English</td>
</tr>
<tr>
<td>ESL</td>
<td>English as a Second Language</td>
</tr>
<tr>
<td>HE</td>
<td>Holistic Evaluation</td>
</tr>
<tr>
<td>IELTS</td>
<td>International English Language Testings System</td>
</tr>
<tr>
<td>LEP</td>
<td>Limited English Proficient</td>
</tr>
<tr>
<td>MLAT</td>
<td>Modern Language Aptitude Test</td>
</tr>
<tr>
<td>MTELP</td>
<td>Michigan Test of English Language Proficiency</td>
</tr>
<tr>
<td>NRT</td>
<td>Norm-referenced Test</td>
</tr>
<tr>
<td>OPI</td>
<td>Oral Proficiency Interview</td>
</tr>
<tr>
<td>SCSA</td>
<td>Standardised Change Score Analysis</td>
</tr>
<tr>
<td>SLA</td>
<td>Second Language Acquisition</td>
</tr>
<tr>
<td>SLEP</td>
<td>Secondary Level English Proficiency Test</td>
</tr>
<tr>
<td>SCSA</td>
<td>Standardized Change Score Analysis</td>
</tr>
<tr>
<td>TEFL</td>
<td>Teaching English as a Foreign Language</td>
</tr>
<tr>
<td>TOEFL</td>
<td>Test of English as a Foreign Language</td>
</tr>
<tr>
<td>TOEIC</td>
<td>Test of English for International Communication</td>
</tr>
</tbody>
</table>