INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

The Role of Handwriting in ESL Writing Assessment

Kevin A. Stanley

A Thesis

in

the TESL Centre

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Arts at

Concordia University

Montreal, Quebec, Canada

December 1999

Abstract

This study explored handwriting bias in ESL writing assessment and the phenomenon of national handwriting styles. It examined the effects of different styles and levels of quality of handwriting on the holistic assessment of intermediate-level ESL writing and how those effects related to the graders' length and depth of experience, as well as their views on writing assessment criteria. The study was also designed to investigate whether the graders could guess the first language (L1) of the writers, using handwriting as a cue.

No bias against "poor" handwriting was found, nor was there a relationship between the graders' length of experience or their views of handwriting as a writing assessment criterion and the holistic grades assigned to writing samples copied in "good" and "poor" handwriting. Explanations offered for the absence of bias in this study include insufficient differences in the degree of legibility between the "good" and "poor" handwriting as well as the graders' familiarity with a range of national handwriting styles.

On the other hand, in the L1 identification, many graders' guesses appeared to have been guided by handwriting cues, rather than textual features. This suggests that if the handwriting corresponds to a national style, this may trigger the application of a set of stereotypes that may in turn influence the graders' responses to their students' writing.

Acknowledgments

First of all, I would like to express my gratitude to my thesis supervisor, Dr. Joanna White. Through all the stages of this thesis, her insight, patience, guidance, encouragement, and enthusiasm kept me on track and made it all seem possible. The amount one normally learns when preparing a thesis was, for me, multiplied by having her as supervisor.

I am indebted to Randall Halter at the TESL centre for his time, advice, and patience with the statistical analyses. He made the abstract concrete for me, and I learned much more than I thought I needed to know.

I am grateful to Professor Patsy Lightbown and Dr. Lori Morris for their support and quick feedback throughout the process. Their comments and suggestions helped guide and expand my research.

The study could not have been conducted without the willing and generous gift of time from the many participants at all phases. They did the tasks asked of them with diligence and professionalism. So many of them were eager to provide comments and insights that helped me clarify the muddle of data and focus on the real issues.

Last, but certainly not least, this thesis would not have been possible without the love, support, and infinite patience of Yuki, who put up with my resistance, procrastination, complaints, and ill-humour for two years. It's over now.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

In first language (L1) education, there has been a great deal of research into a wide variety of bias factors, from gender and race to "halo" effects and the order in which exams are graded. The multicultural schools of North America have responded to the diversity in the classroom by working to eliminate bias from educational assessment.

Bias occurs when factors other than the specific criteria for evaluation influence the grading. Bias can arise from prejudice, but it can often be more subtle. For example, in an L1 high school essay task meant to evaluate a learner's grasp of Canadian history, the grader might be expected to focus on the quality of the content, in other words, whether or not the information is correct, complete, and fully explained. However, if the grader is distracted by the number of spelling errors, the elegance of the style, or the full colour photos printed on the cover page, the specified criteria for assessment are not being followed. In L1 writing assessment, non–content bias factors have been the focus of a great deal of research. The effects of order of reading, length, grader expectations, and penmanship have been extensively studied and debated since early in the twentieth century.

However, in second language (L2) teaching, the above mentioned non-content bias factors have rarely been studied or discussed. As L2 education becomes more specialized, with more and more courses with academic or specific purposes, the risk of confounding the stated evaluation criteria with irrelevant bias factors increases. The cumulative effects of such bias could unfairly retard one learner's educational progress, or give opportunities and rewards to another. The field of L2 teaching is becoming more professional, and teachers are recognizing their responsibility to work towards fairness

and accuracy in their assessment of learners' proficiency. L2 teachers who work with learners in a cross-cultural milieu must be made aware of potential sources of bias to ensure assessment procedures are reliable.

In L2 teaching, student writing is used for many purposes. Among others, it may be used to assign formative or summative grades in a writing course, to diagnose proficiency problems, to place a learner in an appropriate level, or even to determine whether or not a learner is eligible for university entrance. Normally, the graders use a set of criteria for assessment, which are determined by the purpose of the writing task and the goals of the course. The graders may use a holistic or analytic approach to assessment, and they may or may not use a scoring key. No matter what the purpose or procedure of the writing assessment, there are many ways in which non-content bias factors can influence grading.

One non-content bias factor, handwriting, may have a significant role because of the unique characteristics of ESL learners. In the majority of North American adult ESL classrooms, the learners are from a variety of educational and linguistic backgrounds. They come from cultures where literacy and the approach to reading and writing are markedly different from the Western model that most ESL teachers are familiar with. With their different L1 backgrounds, some of these learners bring with them different national or regional styles of handwriting that do not resemble the Western European styles that North Americans are accustomed to. This phenomenon has been studied very little although there is a great deal of anecdotal evidence associated with it.

In order to explore handwriting bias in ESL writing assessment and the phenomenon of national handwriting styles, this study examined the effects of different

styles and levels of quality of handwriting on the holistic assessment of intermediate level ESL writing and how those effects related to the graders' length and depth of experience and their views on writing assessment criteria. The study was also designed to investigate whether the graders could predict the first language of writers using handwriting as a cue.

Chapter 2: Review Of the Literature

It is clear from the following review of the literature that a number of factors have to be considered in the design of a study to investigate the effects of handwriting quality and neatness on L2 writing assessment. These factors include: the type of assessment used, reviewed in Section 2.1, the order in which papers are read (Section 2.2), and the relative length of the papers (Section 2.3). In Sections 2.4.1 and 2.4.2, research on the influence of handwriting and neatness on writing assessment is presented in L2 and L1 contexts respectively. The differences in assessment between word-processed and handwritten papers is reviewed in section 2.5. Research investigating how successfully readers can guess identity information or the L1 of a writer based on his or her handwriting is discussed in Section 2.6. Research on L2 intelligibility, which in some ways parallels research on handwriting and neatness, is discussed in Section 2.7. Section 2.8 reviews the literature on the characteristics of experienced and inexperienced teachers. Finally, the importance of the context of assessment is discussed in Section 2.9.

The literature review reflects the fact that most of the research on writing assessment has been done with L1 writers. The few L2 studies that could be found are clearly indicated.

2.1 Type of Assessment: Holistic vs. Analytic

Second language writing assessment is conducted in a number of ways, depending on the approach of individual institutions and teachers. Some of the research on how assessment is carried out in ESL contexts is reviewed in this section.

Perkins (1983) outlines a number of approaches to L2 writing assessment which fall along a continuum ranging from purely analytic to purely holistic in nature. Analytic measures attempt to objectify the assessment of writing by identifying, classifying, and counting errors or by quantifying the complexity and sophistication of the writing through techniques such as T-unit analysis. While they successfully eliminate many sources of non-content bias, such techniques are time-consuming and labour intensive, and thus are often viewed as impractical for every day classroom use (Brosell, 1986).

Also at the analytic end of the continuum are "indirect" measures of writing, in which learners correct the errors in a text or answer multiple choice questions. The validity of these measures has been criticized on the basis that they do not assess actual writing ability since they ask the learner only to identify correct forms but not to produce them (McColly, 1970; Charney, 1984).

At the other end of the continuum is holistic scoring, which relies on the knowledge of the grader about the teaching goals and performance criteria of the writing task. In holistic grading, a teacher may simply assign a score after reading a student's work, or the teacher may employ a holistic scoring key or guide. Keys usually provide descriptors of the goals and criteria, and when they are used, the grader (often the writer's teacher) roughly matches the writing sample to the descriptor. When no key is used, the grader makes use of some sort of internal criteria. Figure 2.1 shows an example of a scoring grid from Diederich's holistic scoring key. The grid is normally accompanied by several pages of detailed descriptors for each category. The grader circles a number corresponding to the descriptor for each "quality".

	Quality	Low		Middle		High	
General Merit	Ideas	2	4	6	8	10	
	Organization	2	4	6	8	10	
	Wording	1	2	3	4	5	
	Flavor	1	2	3	4	5	
Mechanics	Usage	1	2	3	4	5	
	Punctuation	1	2	3	4	5	
	Spelling	1	2	3	4	5	
	Handwriting	1	2	3	4	5	
					TOTAL:		

Figure 2.1. Diederich's holistic scoring grid (Diederich, 1974).

A common criticism of holistic scoring is that it is unreliable because the vagueness of the descriptors allows too much idiosyncratic variance between scorers (Perkins, 1983; Huot, 1990). Furthermore, the scorers may not be doing what they say they are doing. In an L1 study of writing assessment grading criteria, Freedman (1979) found than many teachers who reported that their principal grading criteria were organization and content were in fact giving feedback about and deducting points mainly for mechanics.

Scoring instruments such as the ESL Composition Profile (Jacobs et al., 1981) were created to improve the reliability of holistic scoring by attempting to quantify the relatively qualitative judgements of holistic scoring. By breaking the criteria down into categories, this instrument tries to balance the weight of various writing skills so that poor performance in one category does not skew the entire grade by distracting the grader. It also separates language development factors, such as vocabulary and grammar, from writing skill factors such as organization and development of ideas. However, it still calls for holistic judgements to be made within the categories.

Using objective measures such as error counts and T-unit analysis, Homburg (1984) investigated the relationship between analytic and holistic assessment, by examining texts that were holistically scored by trained, experienced graders. Homburg suggests that his research validates holistic assessment because the objective measures accounted for 84% of the variance in the holistic scores of intermediate ESL writing.

While the reliability and validity of holistic scoring remain controversial, most L2 writing assessment in classrooms is probably done more holistically than analytically (Carlson and Bridgeman, 1986), and it is clear that holistic scoring allows a number of non-content variables, such as ordering, length, expectancy, and neatness and handwriting to affect the assessment. Any research into holistic writing assessment must address and account for these variables, yet no studies on their effects in L2 assessment could be found. The following sections review the literature on these variables in L1 writing assessment research.

2.2 Research on Ordering Effects

The order in which pieces of writing are read and assessed can have an effect on the scores they receive. Hughes et al. (1983) investigated these "context effects" on essays written by L1 high school students. Fifteen "good", "poor", and "average" essays (as rated by a group of experienced high school teachers) were positioned differently in assessment packets that were given to two other groups of experienced teachers. One of two groups of teachers was given explicit training on context effects and was instructed to read all the essays thoroughly before assigning any grades. The researchers found that in the group that did not receive training, essays that were read immediately following the reading of a poor quality one tended to get better marks than when the same essays are read right after a high quality one. The marks assigned to high quality essays were exaggeratedly high when placed in the context of lower quality essays, and vice versa. The same results were found in the scores of the group that received specific training and instructions designed to minimize the effects of ordering. Hughes and Keeling (1984) found that using model essays or scoring keys did not significantly reduce the effects of context.

2.3 Research on Length

The length or perceived length of a student's piece of writing can have an effect on the grade it receives.

Klein and Hart (1968), investigated a number of non-content variables with 80 law student papers assessed by a group of 17 law professors. They found a strong positive correlation between length (the number of words) and grades assigned. Massey (1983)

also found that length was positively correlated with scores in General Certificate of Education (GCE) 'A' level English Literature essays written by 16-year-old British students.

Peterson and Lou (1991) directly studied the effects of length on the scoring of word processed and handwritten L1 high school essays assessed by teachers. They found that longer papers nearly always received higher scores than shorter papers.

Arnold et al. (1990), comparing the scoring of word-processed and neatly handwritten college-level essays, reported that the handwritten essays had a grading advantage over word-processed versions of the same essays. The researchers attribute some the variance to the graders' reporting that the handwritten versions seemed longer than the word-processed essays, which were printed in a small font.

2.4 Research on Handwriting and Neatness

2.4.1 Handwriting and Neatness in L2

Many experienced ESL teachers perceive distinctive characteristics in the "typical" handwriting styles of students from different countries or regions. For example, a teacher may be able to say that a student's handwriting "looks" Chinese or Latin American. Other than references to the penmanship of Arabic native speakers, no research could be found that investigates this phenomenon, nor could any be found that investigates the impact of handwriting quality on L2 writing assessment.

Many L2 teachers in multicultural adult classrooms report that handwriting quality varies widely among their students and they believe that handwriting seems to present more difficulties to students of certain L1 or cultural groups than others (Neverez et al.,

1979). Odlin (1989) predicts that learners who are literate in syllabic and ideographic systems of writing will have more difficulties learning to write in English than those who already use an alphabetic system, and literate speakers of languages that use the Roman alphabet benefit from positive linguistic transfer.

In the ESL literature, one L1 is often mentioned when discussing difficulties with handwriting: Arabic. Thompson-Panos and Thomas-Ruzic (1983) suggest that adult Arabic L1 students often find handwriting in English to be particularly challenging because left-to-right writing and the Roman alphabet are only learned after childhood for the majority of them. Farquharson (1988) and Santos and Suleiman (1993) point out that the Arabic alphabetic system requires letters to take different forms depending on their position within the word. Also discussing the L2 handwriting difficulties of Arabic speakers, Ball (1986) asserts that difficulties in letter formation, frequent transposition of letters, and problems with "staying on the lines", stem from the strong influence of the right-to-left script of the L1. Learners whose first language uses a left-to-right Roman alphabet would not face the same difficulties, nor would Mandarin and Japanese speakers who usually learn the Roman alphabet in primary school, and are therefore accustomed to left-to-right script. However, the writers cited above offer no research evidence to support their explanations as to why many Arabic speakers have difficulties with Roman script.

As there are differences among the quality and legibility of L2 student writing, it follows that measures of language proficiency that rely on handwritten samples from students may be subject to bias against those for whom clear handwriting is a challenge. Scoring keys such as the ESL Composition Profile (Jacobs et al., 1981) acknowledge the bias and place handwriting in the "mechanics" section, assigning only a few percentage

points to it. However, the research in L1 writing assessment that is outlined below suggests that illegible handwriting could cause the whole essay to be downgraded because of bias. In other words, poor handwriting may influence the evaluation of more heavily weighted categories such as language development and writing skill.

2.4.2 Handwriting and Neatness in L1

Throughout most of the twentieth century, there has been a great deal of research on the question of how handwriting and neatness affect the assessment of writing in L1 educational settings. Nearly all of the studies conclude that pieces of writing with "poor" penmanship get lower marks than equivalent pieces that have "good" penmanship.

An interesting evolution in the purpose of handwriting research has taken place over the years. Until the 1960s, the researchers who identified the bias generally called for improved techniques for handwriting instruction in order to give poor handwriters better skills and thereby level the playing field. In the late 1960s and 1970s, research on handwriting was conducted in the context of a great deal of social science research into bias and discrimination on the basis of race, gender, and other factors. The research findings in this period supported the corresponding de-emphasis of penmanship as a primary educational goal. In the 1980s, the advent of the word-processor inspired a number of new studies on handwriting bias, to determine whether or not word-processed work was evaluated differently from handwritten work. Researchers often suggested that the use of word-processors could eliminate bias in ways that improved handwriting instruction never could.

Most of the studies measuring handwriting bias use a basic design similar to the one represented in Table 2.1. This is a basic 2 X 2 design with two modes of handwriting

combined with two texts, which may be authentic or contrived. This design results in four writing samples that are assessed by two experimental groups.

Table 2.1

Typical 2 X 2 Experimental Design

	Handwriting Mode	
	Good	Poor
Text A	A-Good	A-Poor
Text B	B-Good	B-Poor

Group	Contents Of Assessment Packet
1	A-Good + B-Poor
2	B-Good + A-Poor

In many studies, the same basic design is expanded to 3 X 2, 3 X 3, 4 X 4, and even 10 X 10 and could be used with as many texts and handwriting modes as are practical for experimental purposes.

One typical study was undertaken by Chase (1968), who added a third and fourth factor to create a 2 X 2 X 2 X 2 design. He fabricated L1 high school student answers to essay questions and measured the effects of a number of variables on assessment, including the number of spelling errors, the order of reading, the use or not of a scoring key, and the quality of the handwriting. In this study, only two essays were used, both of which were copied into "good" and "poor" handwriting. The "good" handwriting

received a score of 90 on an instrument called the Ayres Handwriting Scale, while the "poor" penmanship scored 20.

Chase not only found a significant grading advantage for "good" handwriting, but ordering effects similar to those found by Hughes et al. (1983). Poor handwriting was downgraded significantly more when it was read following good handwriting, while good handwriting was upgraded more when it followed bad. Chase also found that more points were awarded to both essays when a scoring key was used. He speculates that without a key, some content features of the essays were overlooked because of positive or negative appearance bias, but attention was drawn to those features by the key.

In another L1 study with fabricated essays, Chase (1979) examined the relationship between handwriting quality and performance expectancy. His graders were given samples of writing accompanied by fictitious transcripts of high school marks. Based on the transcript, the graders characterized the achievement of the "student" as very good, satisfactory or weak, and then evaluated the "student's" writing. The relationship between characterization of achievement (expectancy) and writing scores was significant, but handwriting alone was not unless it was combined with expectancy. When the handwriting was somewhat illegible, high expectancy filled in the gaps for the readers. In other words, a student with straight As could get away with poor handwriting, but a lower-achieving student with poor handwriting was judged much more harshly. However, the expectancy effect was diminished when the handwriting was of good quality.

In a later study, Chase (1986) examined the interaction of handwriting and expectancy with the variables of the graders' knowledge of the sex, ethnicity, and past

performance of the writer. Fabricated essays in good and poor handwriting were attached to contrived report cards and photos of the "writers". Again, he found that poor past performance tended to bias graders against poor handwriters and good past performance had the opposite effect. Chase found that the variables of sex and ethnicity did not have effects by themselves, but that they interacted with each other and with handwriting to generate subtle bias in grading. He also found that graders tended to favour writers who were from the same ethnic groups as themselves.

A number of other studies also found significant differences in the scores given to essays with good or poor handwriting in a variety of L1 educational settings. Briggs (1970) investigated the grades given to 10 authentic compositions written by 11-year-old children in Britain copied in 10 varieties of handwriting. The quality of the 10 handwriting styles was ranked from best to worst by an independent group of teachers. This 10 X 10 design resulted in 100 compositions that were assessed by 10 groups of elementary school teachers, who were asked to rank the compositions, and then to assign a holistic grade. Briggs found that handwriting quality positively correlated with the rankings and holistic grades assigned. He also found that handwriting quality even affects teachers' predictions about a student's future academic success. Using a similar design, Briggs (1980) found that the quality of the handwriting could make the difference between passing and failing standard GCE exams in Britain.

Huck and Bounds (1972) investigated the relationship between the grades given to L1 university essays, and the neatness and clarity of the graders' own handwriting. Using a 2 X 2 design, the researchers fabricated two essays, one slightly more accurate and sophisticated than the other. A panel of three judges selected a "neat", an "average",

and a "messy" (though legible) copier. The average handwriter copied the better essay, which served as a control in the experiment. The inferior essay was copied by both the neat and the messy handwriter. The same panel of handwriting judges grouped the participant graders into the same categories of neat, average, and messy, based on previously obtained handwriting samples. The graders received a packet of two essays, the first being the better essay copied in average handwriting, and the second being the inferior essay copied by either the neat or the messy handwriter. The researchers found that the graders whose handwriting was neat judged the essay copied in poor handwriting significantly lower than when it was neatly copied. The graders whose handwriting was messy did not differentiate significantly between the two.

Marshall (1972) used a 3 X 3 design to examine the interaction of the number of spelling errors and the quality of handwriting in a modified L1 high school essay. Sixteen versions of the essay were prepared, combining four levels of spelling errors with four modes of transcription: typed, "neat" handwriting, "fair" handwriting, and "poor" handwriting. The researcher did not report how the quality of the handwriting was assessed. A scoring key, with instructions to focus only on content was included with the essays, which were assessed by 480 high school teachers. The researcher found that, even with the scoring key and instructions, the neatly handwritten essays received significantly higher grades than the poorly written versions. Marshall also found that the graders frequently overlooked the spelling errors in all the handwritten versions while they focussed upon or "noticed" the same errors in typed versions of the essays.

Soloff (1973) used model L1 junior high school history essays from a study guide, which were copied by hand by two students. According to the researcher, one of the

resulting papers was "neat", and the other was "sloppy". Along with legibility, Soloff included the number of crossed out words and false starts in her criteria for determining "sloppiness". In this 2 X 2 design, the two groups of junior high school teachers who assessed the essays were significantly influenced by the neatness of the essays, though they were instructed to judge the content alone. The researcher interviewed the participants after the experiment, and many reported that they believed that "sloppy" work was a sign of poor general writing skills.

Markham (1976) used a 9 X 9 design, combining three levels of quality for content and three levels of quality for handwriting in descriptive paragraphs from L1 fifth grade writers. The quality of the content and handwriting in the papers was judged by an independent group. The graders were in-service and student teachers. Markham found that paragraphs with good handwriting consistently outscored those with poor handwriting. Using data from a questionnaire completed by the participants, the researcher found that the age, length of experience, and education level of the graders had no significant effect on the scores they assigned.

Sloan and McGinnis (1982) started with 500 short essays written by ninth graders, which were graded by experienced graders using the Diederich holistic scoring method key (Diederich, 1974). A group of experts in the "Palmer Handwriting Method" rated the same essays in terms of handwriting quality. A subset of 45 of the papers was randomly selected, 15 from the lower third, 15 from the middle third, and 15 from the upper third in handwriting quality. Each of these papers was carefully copied by the same handwriting experts, and resubmitted to the holistic scorers. This time, the expertly copied papers all

received significantly higher grades than the originals, though the greatest difference was found in the upper two-thirds.

Sprouse and Webb (1994) found that even L1 elementary school spelling tests were subject to appearance bias. In a 2 X 2 design, teachers were asked to grade two spelling tests of 20 words, one in poor handwriting, the other in good handwriting. Both tests contained the same five spelling errors. The mean number of errors found in the test in good handwriting was 4.4, while the mean for the poor handwriting was 6.4. In other words, errors that weren't there were found in the poor handwriting, but others were overlooked in the good handwriting.

The only study that could be found that had contrary findings to those outlined above was conducted Massey (1983). The researcher did a post hoc analysis of the grades given to essays written in English Literature for British standard GCE 'A' level exams at the high school level. One of the variables, neatness, as judged by the researcher, had no significant impact on the scores assigned.

To sum up the research on appearance bias in L1 writing assessment, apart from Massey (1983), all of the studies consistently find varying degrees of bias against poor handwriting. Most of the studies use fabricated or modified texts that are copied in different levels of handwriting quality. To operationalize handwriting quality, the researchers used panels of judges, rating scales, handwriting experts, or personal opinion.

2.5 Handwriting vs. Word Processing

With the arrival of word-processing in the late 1980s, several researchers studied its impact on the assessment of writing. There was concern that writers who had access to

word processors may have an advantage over those who did not, and a debate ensued over whether or not word-processing skills should be integrated into writing curricula (Powers et al, 1994; Arnold et al, 1990).

Arnold et al. (1990) noticed that in college-level L1 composition courses, word processed papers got significantly better grades. They tried to determine whether it was the editing and revision advantages of word processors or a grader preference for reading printed text that gave word-processed papers the edge. The researchers randomly selected several hundred handwritten essays from past examinations and typed them, verbatim, using word processors. Then, with handwritten originals mixed with word-processed versions, two groups of experienced graders assessed them in a 2 X 2 design. The quality or legibility of the handwriting was not a design factor. To their surprise, the researchers found that the handwritten originals received higher scores. In post-experiment interviews, the graders reported a preference for reading the handwritten essays, though they were more difficult to read. The graders also perceived the handwritten papers to be longer (as discussed in section 2.3) than the word-processed versions (which were printed, single spaced, by a dot matrix printer), and reported having higher expectations of word-processed work. Some graders said that they judged spelling or typographical errors more harshly in word-processing, and tended to "fill in the gaps" or give the benefit of the doubt in handwriting. The researchers hypothesize what they called the R-E-A-D effect: Reader Empathy Assessment Discrepancy, which appears to favour handwritten papers. Scorers reported that they felt "closer" to the writer of a handwritten text, and that somehow the "voice" of the student was absent in the word-processed versions.

In a similar study with different results, Peterson and Lou (1991) had word-processed and handwritten versions of L1 grade 9 students' papers scored, and they found no significant differences between the two modes in terms of scoring. However, two findings were similar to those of Arnold et al. (1990). First, as mentioned in section 2.3, they found that longer papers nearly always received higher scores, whether they were handwritten or typed. Second, the scorers reported that errors of spelling, punctuation, and capitalization "jumped out" at them in the word-processed papers, implying that the same errors may have been overlooked in handwritten versions. The quality of the handwriting was not a variable in their study.

Powers et al. (1994) found handwritten college-level L1 essays fared better than word-processed ones. The researchers took a handwritten essay and a different word processed essay from each of 32 student writers. The original handwritten essays were then word processed, and the word-processed essays were handwritten by several transcribers, who attempted to emulate the actual handwriting of the original writers. The transcribers also introduced crossed-out words and erasures to make the copies seem authentic. The scores given by a small group of trained and experienced holistic graders showed a significant advantage for the handwritten versions.

Powers et al. speculate that in word-processed form, the papers seemed shorter and surface errors were more salient to scorers, which supports the conclusions of Arnold et al. (1990) and Peterson and Lou (1991). They also suggest that handwritten papers showed obvious signs of revision (crossed-out and partially erased words) that are not apparent on a neat-looking word-processed version. This may suggest the amount of care taken or effort made by the writer, generating sympathy or appreciation on the part of

scorers. In later phases of the study, the researchers tried to compensate for the effect of perceived length by using larger point sizes for the fonts and double-spacing in the word-processed versions, and they conducted special training sessions to discuss bias with the scorers. These efforts reduced the effect, but some bias in favour of handwritten work remained.

In another L1 college-level study that involved training the graders, Sweedler-Brown (1992) compared the scoring of very good handwriting, poor handwriting and word processing, and measured the effect of scorer training on the grades. In this 3 X 2 X 2 design, half of the graders were specifically trained with seminars on appearance bias. The quality of the handwriting was determined by Diederich's criteria, which are shown in Table 2.2.

Table 2.2

Diederich's Descriptors for Handwriting and Neatness (Diederich, 1974)

Level	Descriptor
High	The handwriting is clear, attractive, and well spaced, and the rules of manuscript form have been observed.
Middle	The handwriting is average in legibility and attractiveness. There may be a few violations of rules for manuscript form if there is evidence of some care for the appearance of the page.
Low	The paper is sloppy in appearance and difficult to read. It may be excellent in other respects and still get a low rating on this quality.

Sweedler-Brown found good handwriting significantly outscored both word-processing and poor handwriting, and there was no difference between the scores for word-processing and poor handwriting. She also found that training had no significant effect. Sweedler-Brown speculates that because some of the graders who were specially trained were already very experienced graders, they may have been more resistant to training. On the other hand, some of the untrained graders were also relatively inexperienced, and given the opportunity, they may have been more open to training to overcome "appearance bias" than the experienced graders. Since the experienced and less experienced graders were evenly mixed between the groups, the effects of the training may have been reduced.

In general, legibly handwritten essays get better marks than word processed counterparts. However, it should be noted that all of this research on the differences in assessment for handwriting and word-processing was conducted in the late 1980s and early 1990s, when computer use among students and teachers was not nearly as widespread or as familiar as it is today. Printer technology and word-processing software have also changed, rendering word-processed work far more readable and attractive than a decade ago when these studies were done.

2.6 Determining Writers' Identities from Handwriting

There has been some research on how successfully readers can guess the gender or ethnicity of writers, using cues in the texts, including handwriting.

Loewenthal (1980) and Eames and Loewenthal (1990) claim that native speakers can correctly guess the gender of other native-speaker handwriters who copy content-neutral samples of writing 75% of the time.

Emerling (1991) tested whether or not readers could guess the ethnic background (white, Hispanic, African-American, or Asian) and gender of native-speaker writers of college compositions. Three readers guessed the gender of 194 writers successfully 87.6% of the time, and correctly identified white writers 94% of the time, Asians 53%, African-Americans 32%, and Hispanics 12%. However, the researcher speculates that when the guesses were successful, the content of the compositions probably contained clues about the writers' identities.

Sprouse and Webb (1994), asked teachers to guess the gender of the pairs of students whose spelling tests they were grading. All of the participants attributed the illegible tests to male students, whether they were written by males or not.

Other than comments by Thompson-Panos and Thomas-Ruzic (1983) and Ball (1986) on the distinctiveness of certain Arabic speakers' English handwriting, there appears to be no research on whether or not graders can determine the first language of an L2 writer from his or her handwriting.

2.7 Research on Intelligibility

While no research could be found on handwriting in L2 writing assessment, one area of L2 research that may be related is the study of intelligibility. Poor handwriting could be considered similar to a strong "accent", in that it interferes with the reader's ease of comprehension. A reader has to expend extra effort to decipher writing that is difficult to read, perhaps in the same way a listener has to work harder to interpret heavily accented speech. In the latter case, studies have investigated the extent to which there may be affective results (i.e. irritation) caused by the extra work the listener has to do.

Brennan and Brennan (1981) is typical of a number of studies measuring native and non-native speaker attitudes towards accented speech. Two groups of high school students, one consisting of Anglos and the other made up of Mexican-Americans, listened to tape recordings of Mexican-Americans with various degrees of accented speech in English. The participants were asked to rate the speakers in terms of socio-economic status and the degree of solidarity they felt with the individual. The researchers found that, for both groups, as the degree of accentedness increased, ratings of status and solidarity decreased.

Listener familiarity with non-native speaker accents is an important factor in many intelligibility studies. The more familiar or experienced the listener is with accents in general, or with a specific accent, the higher the listener rates comprehensibility (Gass and Varonis, 1984; Munro and Derwing, 1995; Derwing and Munro, 1997). Gass and Varonis (1984) also noted that more experienced listeners had a greater ability to distinguish grammatical errors from pronunciation errors. This finding may have implications for a study on the effects of the "typical" handwriting styles of certain L1s on writing assessment. The degree of familiarity a teacher has with different handwriting styles may affect the teacher's reaction to them.

However, there is an interesting contrast between research on intelligibility and research on handwriting. Varonis and Gass (1982) found that increasing the grammatical errors in recorded speech caused exaggerated perceptions of the strength of an accent, while Peterson and Lou (1991) and Powers et al. (1994) found that legible handwriting (as opposed to word-processing) may cause such errors to be overlooked or ignored.

2.8 Experience of Graders

Very little research has investigated the different approaches to assessment taken by experienced and inexperienced graders. The study by Sweedler-Brown (1992) discussed in Section 2.5 made note of the levels of experience of the graders and speculated on their receptiveness to training. In a study that surveyed the approaches taken by experienced and inexperienced evaluators of ESL writing, Carney (1973) reported that less experienced graders tended to focus more on mechanics and "formed broad impressions of student texts", while more experienced graders made more uniform judgements and were more likely to use a hierarchical approach to assessment, starting with organization and content, before moving on to rhetorical structure, and finally looking at surface errors.

In support of these findings, Cumming (1990) recorded concurrent think-aloud reports from a group of experienced ESL teachers and a group of novice student-teachers who holistically assessed ESL compositions in terms of "language use", "substantive content", and "rhetorical organization." He then analyzed the decision-making strategies and assessment behaviours of the two groups and found that experienced teachers used sets of strategies and behaviours to evaluate writing and linguistic skills that were systematic, complex, slightly idiosyncratic, but highly consistent and reliable. During the think-alouds, they usually rationalized their decisions based on clear criteria. On the other hand, the student-teachers tended to verbally edit the texts, focusing on surface errors, before finally arriving at an impression of overall quality without giving any explicit criteria. For the "language use" ratings of the papers, the novices consistently gave higher marks than the experts, while for "substantive content" and "rhetorical organization", the

student teachers had low interrater reliability and the experienced teachers were very consistent.

2.9 Context of Assessment

Virtually all of the studies reviewed above try to isolate bias variables by removing the writing assessment task from its real-world context. Assessment that is used for level placement or to screen university applicants may or may not conceal identity information, such as sex, L1, or ethnicity, from the grader. Briggs (1970, 1980) points out, however, that real classroom evaluation is rarely done without the teacher knowing exactly who the writer is, and that many factors may influence a teacher's grading. For example, a teacher may grade up for encouragement, or as a reward for good behaviour, or grade down because of the work-habits, personality or attitude of the student. Briggs' point is that typical assessment procedures are extremely complex and involve the interaction of dozens of variables, and that findings based on the isolation a single variable must be applied with caution to real-world situations. Diederich (1974) observes that a teacher's knowledge of the student's abilities and past performance strongly influences the grades given. Diederich's point is supported by Chase's (1979) study on handwriting and expectancy.

2.9 Research Questions

Given the wide range of research that has been done on the effects of handwriting and neatness on L1 writing assessment, and the absence of any such studies in L2, it is not clear how the L1 findings might apply to an adult L2 context. ESL teachers' reactions and judgments to handwriting and neatness may be tempered by their varying degrees of experience with different handwriting styles produced by students with different linguistic and ethnic backgrounds. A teacher's approach to L2 learning and the relative importance she or he assigns different writing assessment criteria may also influence those reactions. The current study was designed to address these issues.

The research questions are the following:

1. What effect does handwriting quality and/or neatness have on the holistic assessment of L2 written work?

2. How does this effect relate to the length and range of teachers' experience?

3. How does this effect relate to teachers' views on writing assessment criteria?

4. Can ESL teachers accurately guess the first languages of writers when handwriting is used as a cue?

Chapter 3: Methodology

3.0 Overview of the Study

3.0.1 Four Stages

In this study, writing samples were fabricated and handwriting samples were collected from ESL learners. Both sets of samples were assessed by a group of ten ESL teachers. This group also tried to guess the L1 of the supposed writers of the fabricated samples. The writing samples that elicited both a high level of agreement in scoring among the participants and a low level of agreement on the guesses of the L1 were selected. The handwriting samples that represented the middle range in terms of quality were also selected.

The selected handwriters copied the selected writing texts in their own hand. These samples were photocopied and arranged in packets for four different groups of ten ESL teachers to assess. The results of the latter groups' assessment were analysed for handwriting bias and the ability to identify the L1. Information obtained from a questionnaire was also analysed.

This study had four stages: 1) the fabrication of the original writing samples; 2) the assessment and selection of the writing samples and handwriting styles to use in the main experiment; 3) the copying of the typed writing samples into handwritten form; and 4) the main experiment in which the handwritten samples were assessed. The four stages are presented graphically in Figure 3.1.

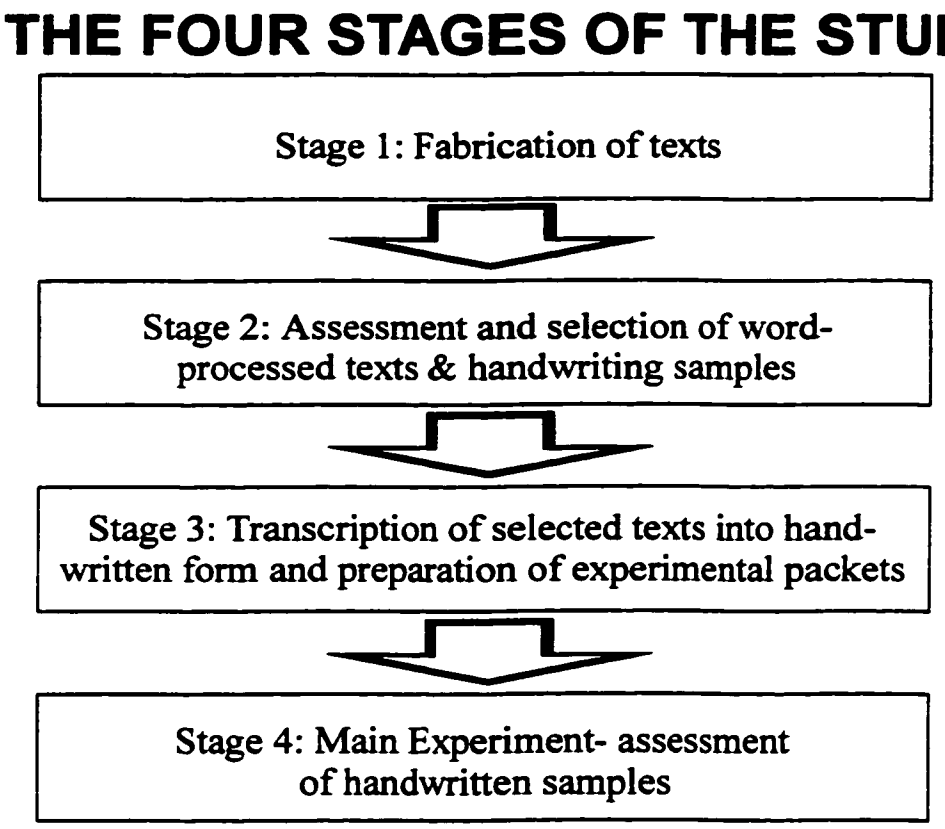

Figure 3.1. The four stages of the study.

3.0.2 Participants

There were three groups of participants in this study. The first group consisted of ten university-level credit ESL teachers who assessed word-processed texts to aid in the selection of the samples that were copied into handwritten form. The same group judged handwriting quality to assist in the selection of the handwriters who did the copying. Hereafter, this group of participants will be referred to as the "selection group".

The second group consisted of ten beginner-level learners studying in a non-credit intensive ESL course. They came from a variety of L1 backgrounds. It was this group's handwriting that was judged by the selection group. Six of these students copied the

word-processed texts by hand for the main experiment in Stage 4. These six are referred to hereafter as the "copiers". The selection of the copiers is discussed in Section 3.2.

The third group of 40 participants were instructors at a large non-credit intensive English for Academic Purposes (EAP) program at a Canadian university. This program has a particular institutional culture and lengthy training process that results in a coherent program with teachers who tend to have similar views on assessment and use very similar methodological approaches to teaching. The length of teaching experience in multilingual adult ESL classrooms of the instructors ranged from three years to more than 25, with an average of 8.96 years. Further data about the participants appears in Table 3.1. This group of participants will be referred to hereafter as the "main group".

Table 3.1

Main Experiment Participant Data

		N
Age	20-29	5
	30-39	9
	40-49	13
	50-59	13
Sex	F	29
	M	11

A frequent classroom task at the intermediate levels at this institution is the "written reconstruction". This is the final step in a listening activity in which learners individually reconstruct a narrative audio or video text in writing, after several group-centered tasks involving comprehension questions and note-taking. The writing step is usually intended as a written "recycling" of the content, and as a vehicle to develop writing competence, but it is occasionally used for diagnostic purposes. When teachers at this institution assess such samples of writing, they usually do so quickly and holistically.

3.1 Stage 1: Fabrication of the Texts

3.1.1 Creating the Texts

The first stage in the study was to create writing texts. The decision to use fabricated, rather than authentic, texts was made after reviewing many L1 writing assessment studies. Using fabricated texts allows for control over the number, type, and position of errors, and it eliminates any effect caused by differences in length (Arnold et al., 1990; Klein and Hart, 1968; Peterson and Lou, 1991). The use of fabricated samples also removes the influence of "evidence of revision" found by Soloff (1973), which might signal that greater effort had been made in revising the work.

A final rationale for the choice of fabricated samples concerns the perception of completeness. The third group of participants, the teachers in the intensive program, would be told that the samples were in-class written reconstructions of a listening task. It was anticipated that a bias might occur if the graders perceived that a writer had been more or less successful in completing the writing task. For example, if the task was to reconstruct a listening text in writing, and one writer provided more information than

another, the grader might inadvertently favour the writer of the more "complete" text. Completeness may be related to better listening skills, or simply greater speed or confidence in writing, rather than superior overall writing skills. Therefore, the samples had to be controlled for the thoroughness of the information content, while at the same time they had to seem to be authentic samples of in-class writing.

The first step in the fabrication of the samples was to collect a set of in-class written reconstructions from an intermediate class in the intensive program. The writing was done as the final task in a listening activity. The students had watched an off-air video about a woman who volunteered to work undercover for police to help them arrest a drug dealer. They had worked in groups to answer questions and take notes, then they orally reconstructed the information from the video in pairs. Finally, they wrote the story individually from their notes.

The class set of 16 written reconstructions provided the raw materials for the researcher to create 10 original texts, in which the content of the video was reproduced. The reconstructions served as models for both the type of information that students included and for the types of errors they made. An analysis of the reconstructions produced lists of information units, misspellings, and common errors made by the writers. An error inventory (Burt & Kiparsky, 1972) was consulted to categorize and enumerate the errors.

After the first draft of the samples was completed, a careful analysis of the errors was conducted. Errors were counted and distributed similarly across the samples. That is, all the samples had approximately the same number of errors of the same type occurring

in the same relative position in the text. The types of errors included are shown in Table 3.2.

Table 3.2

Error Types in the Fabricated Samples

Error Type	Example
Spelling of proper nouns, common words and uncommon words	Jimy, secrete, fraightened
Number / subject-verb agreement	One pounds of drugs, her daughter are off drugs
Irregular verbs	They quitted drugs
Omitted copula	Now her daughter in jail
Passives	She was worry about her daughter
Wrong or omitted preposition	She was introduced for the drug dealer
Wrong or omitted determiner	drug dealer's name is Jimmy

These types of errors were chosen because they are typically made by intermediate-level students who speak a variety of L1s. Thus, the L1 identification tasks in later stages of the study would not be unduly influenced by linguistic features of the writing that some may consider typical of a particular L1. All of the error types in Table 3.2 occurred in at least half of the original written reconstructions.

Table 3.3

Length Statistics For Fabricated Texts

Text	Words	Characters	Paragraphs	Lines
1	166	746	4	20
2	170	729	4	19
3	174	748	4	19
4	168	747	4	18
5	172	743	4	19
6	169	744	5	21
7	174	726	5	19
8	170	730	4	18
9	164	746	5	20
10	159	746	5	20
Mean	168.60	740.50	4.40	19.30
SD	4.65	8.57	0.52	0.95

The ten texts were constructed to be approximately the same length. They were word processed and printed in a 14-point Arial font, double spaced, so they appeared to fill 80% to 90% of an 8½ X 11 page. This font size was chosen because word counts of

the authentic samples indicated that 14-point type yielded text that appeared to be approximately the same length as average handwriting, thereby avoiding the length perception problems encountered by Arnold et al. (1990), which were noted in Chapter 2. The number of words varied from 159 to 174, with a mean of 168.6. The number of characters (excluding spaces) ranged from 726 to 748, with a mean of 740.5. The number of lines of text ranged from 18 to 21, with a mean of 19.3. Six texts had four paragraphs, and four had five paragraphs. The length statistics are shown in Table 3.3.

3.1.2 Collection of Handwriting Samples

To identify the handwriters who would be used to copy the fabricated texts, in-class writing samples from a low-level class were collected. From the 20 writers, the researcher selected 10 samples that represented a range of handwriting styles from a variety of L1s, which reflected the population of the school. Two samples were written by Arabic native speakers, two were Korean, two were Latin Americans, two were Mandarin speakers, one spoke Vietnamese and one spoke Japanese. Names were removed and the samples were photocopied.

3.2 Stage 2: Assessment and Selection of Texts and Handwriting

3.2.1 Text and Handwriting Assessment

To select the texts and handwriters that would be used in the main experiment in Stage 4, an independent group of ESL teachers graded the fabricated texts, guessed the L1s of the "writers" and ranked the quality of the ten handwriting samples. The participants in the selection group were university credit ESL teachers, experienced with

international students. Their recent experience was with students at a higher level of written proficiency than that of the "writers" of the fabricated samples; however they all had previous experience with writers of lower proficiency. Five data collection sessions were held for this stage, to accommodate the participants' teaching schedules.

3.2.2 The Packets

Envelopes containing the ten fabricated, word-processed texts were prepared and numbered. The texts were labeled with three-digit numerical codes to avoid any perception of ranking. In other words, they were not number 1 to 10, in case those numbers led any of the participants to perceive them as being presented in a particular order. The contents of each packet were ordered in the same way for all participants.

Each text had a slip of paper stapled to it, on which there was a numerical code to identify the text and the participant number. On each slip there was a series of numbers, from 0.0 to 10, in 0.5 point increments. (see Appendix D-3) On the back of each text, there were blank spaces on which to answer the L1 identification questions (Appendix D-12).

3.2.3 Step 1: Holistic Grading of the Texts

The participants were first verbally instructed that they would be asked to holistically grade ten pieces of writing, which originated in an intermediate class of international students. The task that led to the writing (a written reconstruction of a video listening activity) was described to them. The participants were asked to quickly look through the samples to establish in their minds the approximate level, and not to grade them according to the higher levels they were currently teaching. The participants were

told to imagine the purpose of the grading was to provide feedback to the students about their writing ability at this point in their course.

The participants circled a grade out of ten on the slips that were stapled to each text. These grading slips were chosen to eliminate ambiguity in the grading by forcing all the participants to use the same scale. After this step, which took between 10 and 15 minutes, depending on the individual, the participants were instructed to tear off the slips and put them in a separate envelope, provided for that purpose.

3.2.4 Step 2: L1 Identification

After all the participants had finished the holistic grading, written and oral instructions were given for the L1 identification step. These instructions were not given earlier, to avoid making any suggestions about the true purpose of the experiment in the first step.

The participants were asked to re-read the texts, then make a guess as to the L1 of the writer, indicate the degree of certainty of that guess, and identify which features of the writing led them to make that guess. There were blank spaces on the back of each text on which they could write their responses. If they could not guess the specific L1, they were permitted to write a language group, or failing that, they could write "I don't know." The participants were encouraged to make their best guess. This step took between 15 and 25 minutes to complete, after which the participants returned their texts to their envelopes, which were then collected.

3.2.5 Step 3: Handwriting Quality Assessment

The third step was to rank the quality of the handwriting samples. Each participant got a new packet containing the ten handwriting samples. They were

instructed to ignore grammar and spelling errors, and to focus on the quality of the handwriting. The participants ranked the samples, from best to worst, and then recorded the rankings by listing the sample numbers on a separate form provided for that purpose. This step took around five minutes for most participants.

3.2.6 Participant Comments During Stage 2

Various members of the selection group in this stage made unsolicited comments following the completion of the three steps of Stage 2. Some participants reported a desire to know more about the context of the class when holistically scoring the fabricated samples. Some also wanted clearer criteria for grading, though they were asked to grade holistically.

The L1 identification task was the most challenging for most of the participants. Many reported that their guesses were really just guesses, and they felt they had little to go on. Several remarked that some of the writers made errors just like certain individuals in their current classes, and those provided their clues to the L1. Others reported a greater interest in or experience with contrastive linguistics, and drew on that knowledge. Almost all the participants reported that they were not comfortable or satisfied with the judgements they made.

The handwriting quality ranking was done quickly by the participants. They reported finding it relatively easy to rank the handwriting samples, and seemed very confident in their judgements. One pointed out that she had no training or experience rating handwriting, and that she had never before given it any thought. Another said it would have been easier to identify the L1s of the writers with the handwritten samples.

3.2.7 Text Scores

In the first step of this stage, the participants gave a holistic grade to each text. The grades are recorded in Table 3.4. One participant gave the same grade for each text, so these data were discarded, leaving nine data sets.

Table 3.4

Grades for Texts

	TEXTS									
Participant	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
1	5.5	7.0	6.0	5.0	4.5	6.0	4.5	4.5	5.5	5.5
2	7.5	8.5	9.5	7.5	7.5	7.5	7.5	7.0	7.0	7.5
3	7.0	7.5	8.5	6.5	6.5	7.0	6.0	7.5	6.5	7.0
4	5.5	7.0	6.0	6.5	6.5	7.0	5.5	6.0	5.0	5.5
5	7.0	7.5	9.5	6.5	8.5	7.0	7.0	6.5	9.0	6.5
6	6.0	6.5	7.0	6.0	6.0	6.5	5.5	6.5	6.5	6.0
7	7.5	9.5	9.5	7.5	7.0	8.0	8.0	8.5	8.0	6.0
8	7.0	6.0	7.0	8.5	6.0	9.0	7.5	6.5	8.5	5.5
9	7.0	8.5	8.5	7.5	6.5	8.0	6.0	7.5	7.0	6.0
Mean	6.67	7.56	7.94	6.83	6.56	7.33	6.39	6.72	7.00	6.17
SD	0.79	1.10	1.47	1.03	1.10	0.90	1.17	1.12	1.32	0.71

The purpose of the scoring was to determine which texts generated the highest degree of agreement among the participants in terms in terms of scores. It had been previously decided that the four texts with the highest standard deviations (SD) would be discarded, and the six with the lowest standard deviations would be kept for the next stage of the study. As shown in Table 3.5, the grading of texts #10, #1, #6, #4, #5, and #2 produced the lowest SDs, while the grading of #8, #7, #9, and #3 produced the highest.

Table 3.5

Ranking of Standard Deviations of Text Scores

TEXT	SD	
#10	0.71	
#1	0.77	
#6	0.90	
#4	1.03	LOWER
#5	1.10	
#2	1.10	
#8	1.12	
#7	1.17	
#9	1.32	HIGHER
#3	1.47	

3.2.8 Results of L1 Identification

The L1 Identification task was included to determine whether any of the fabricated texts had features that would suggest to the participants that the writer spoke a particular L1. Some participants wrote vague or multiple answers and reported low levels of certainty about their guesses.

Table 3.6

L1 Identification

Text	French	Spanish	Arabic	Slavic	Asian	Don't Know
#1	2			2	4	2
#2	4	1	1	2	2	
#3	3		1	2	2	2
#4	1	1	1	1	6	
#5		5	3		2	
#6	4	2		1	3	
#7	5	1	1	1		2
#8	5		2	1	1	1
#9	5	1		2	2	
#10	1		1		6	2

As can be seen in Table 3.6, most of the texts generated a wide range of guesses, though no L1 was guessed more than 60% of the time. In the case of the "Asian" and "Slavic" guesses, these categories were created because most of the participants wrote answers such as "Chinese/Japanese" or "Oriental" or "Russian/Eastern European" on certain texts. Therefore, all responses related to East Asian and Slavic languages were collapsed into these two categories.

3.2.9 L1 Identification Levels of Certainty

Many of the participants had relatively low confidence in their L1 guesses, as determined by the ranking they provided from 1 (very certain) to 4 (very uncertain). Table 3.7 shows the level of certainty for all the participants and samples. Blank cells appear where the participant wrote "I don't know."

As shown in Table 3.7, only one participant (4) had consistently high levels of certainty, while the others almost never characterized their guesses as "very certain". The text that led to the highest level of certainty was #5, which had a mean level of 2.20. The mean for all the guesses was 3.02.

Table 3.7

Levels of Certainty in L1 Identification

	Texts										
Participant	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Mean
1		3		4	3	4			4		3.60
2	3	2	2	4	3	4	2	2	2	2	2.60
3	2	3	3	4	2	2	3	4	4	3	3.00
4	1	1	2	1	1	1	1	1	1	2	1.20
5	3	4	4	2	1	3	4	4	3	4	3.20
6	4	4	3	4	2	4	4	4	3	4	3.60
7	4	4	4	4	3	4		4	4	4	3.89
8	3	4	2	3	2	3	3	2	2	4	2.80
9	4	4	4	4	3	4	4	3	4	4	3.80
10		4	2	4	2	2	3	4	2		2.88
Mean	3.00	3.30	2.89	3.40	2.20	3.10	3.00	3.11	2.90	3.38	

The final factor in determining whether the participants in this phase could successfully identify the L1 of the writer was the features of the text that they detected that led them to make their guesses. Again, many of the features mentioned were vague and general. For example, errors with "articles" pointed to an Asian writer for some, but to a Slavic writer for others. Errors with "passives" or "verbals" were used to identify French, Spanish, Arabic, Asian and Slavic. The only text that had any particular feature that repeatedly stood out was #5, which contained the misspelled word "bolunteered". It elicited 5 Spanish, 2 Arabic and 2 Asian guesses, all citing the same error. This error may have led to its relatively high (2.20) certainty rating.

3.2.10 Selecting the Texts

After reviewing the results for the holistic grading, the L1 identification, and the levels of certainty, the following texts were selected for use in the main experiment in Stage 4: #1, #2, #4, #6, #8, and #10. Texts #3, #7, and #9 were rejected because of the high SDs the grading produced, and #5 was rejected because of the spelling error that so many of the participants in the selection group cited as an L1 indicator.

3.2.11 Results of Handwriting Quality Assessment

Table 3.8 shows the numbers corresponding to the ranking of the handwriting samples by the 10 selection group participants. The ranking numbers were averaged to produce an overall ranking of handwriting quality.

Table 3.8

Handwriting Quality Rankings from Best to Worst

	Participants												
HW	P. 1	P. 2	P. 3	P. 4	P. 5	P. 6	P. 7	P. 8	P. 9	P. 10	Mean	SD	Rank
G	1	1	1	4	2	1	4	1	1	2	1.80	1.23	1
E	2	2	2	2	1	4	2	5	2	1	2.30	1.25	2
F	3	4	5	1	4	2	1	3	4	4	3.10	1.37	3
L	4	6	4	3	3	5	3	2	3	3	3.60	1.17	4
T	5	3	3	5	5	3	5	4	5	5	4.30	0.95	5
J	8	5	6	7	6	7	7	7	7	6	6.60	0.84	6
Y	7	8	7	6	10	6	9	6	6	7	7.20	1.40	7
M	6	9	9	9	7	8	8	8	9	8	8.10	0.99	8
B	9	7	10	8	8	9	6	10	10	9	8.60	1.35	9
I	10	10	8	10	9	10	10	9	8	10	9.40	0.84	10

3.2.12 Selecting the Copiers

The extreme high and low rankings, four in all, were rejected to leave the middle six. In other words, handwriters G, E, B, and I were rejected, and handwriters F, L, T, J, Y, and M were retained to copy the selected texts in Stage 3. Table 3.9 shows the L1 of each of the six. The rationale for this selection was that extremely good or extremely poor handwriting might provide clues as to the true nature of the study to the main group of

participants in Stage 4. The six retained handwriters would copy the selected fabricated writing samples in the next step.

Table 3.9

L1 and Ranking of Handwriters

Handwriter	Handwriting Quality Rank	L1
F	1	Mandarin
L	2	Hebrew
T	3	Vietnamese
J	4	Spanish
Y	5	Japanese
M	6	Arabic

3.2.13 Summary of Results of the Selection of Samples

Of the ten fabricated writing texts, three were rejected: #3, #7, and #9, because of their high standard deviations, and #5 because of its one spelling error that elicited both moderate agreement on the L1 and a high level certainty in the guesses. The remaining six were : #1, #2, #4, #6, #8, and #10. The six middle-ranked handwriters (F, L, T, J, Y, and M) were also selected.

3.3 Stage 3: Copying Texts and Preparing Experimental Packets

3.3.1 Copying the Texts by Hand

In this phase, the six handwriters copied the six fabricated texts by hand. The two "better" handwriters, (F and L), and the two "poorer" handwriters (Y and M) were asked to copy four different fabricated texts. Each experimental packet for Stage 4 would contain a different text copied by each of the four handwriters. Handwriters T and J were in the middle of the rankings, so they were asked to copy just one text each, which would be placed in all the experimental packets for Stage 4. These "neutral" samples, written in neither good nor poor handwriting, would serve to measure interrater reliability among the four experimental groups. Figure 3.2 illustrates the copying.

The handwriters were asked to copy the samples verbatim in their normal handwriting. They were given written instructions (see Appendix C) that were orally explained. They were told that final handwritten versions should fill one page, double spaced, and should appear to be done as an in-class writing. They were not told that the study was looking at handwriting in particular, but that it was investigating "how teachers grade writing". This was to avoid any self-conscious modifications of their handwriting.

The researcher verified that all errors had been copied exactly, and that no new errors were introduced. Each of the handwriters inadvertently corrected certain errors in the samples, but these were found by the researcher, and subsequently changed by the copiers to conform with the texts. In the course of carefully copying the texts, one of the copiers' handwriting improved over the original handwriting sample that was judged in Stage 2. This copier was asked to re-copy less carefully, and the second effort conformed more closely to the original handwriting sample.

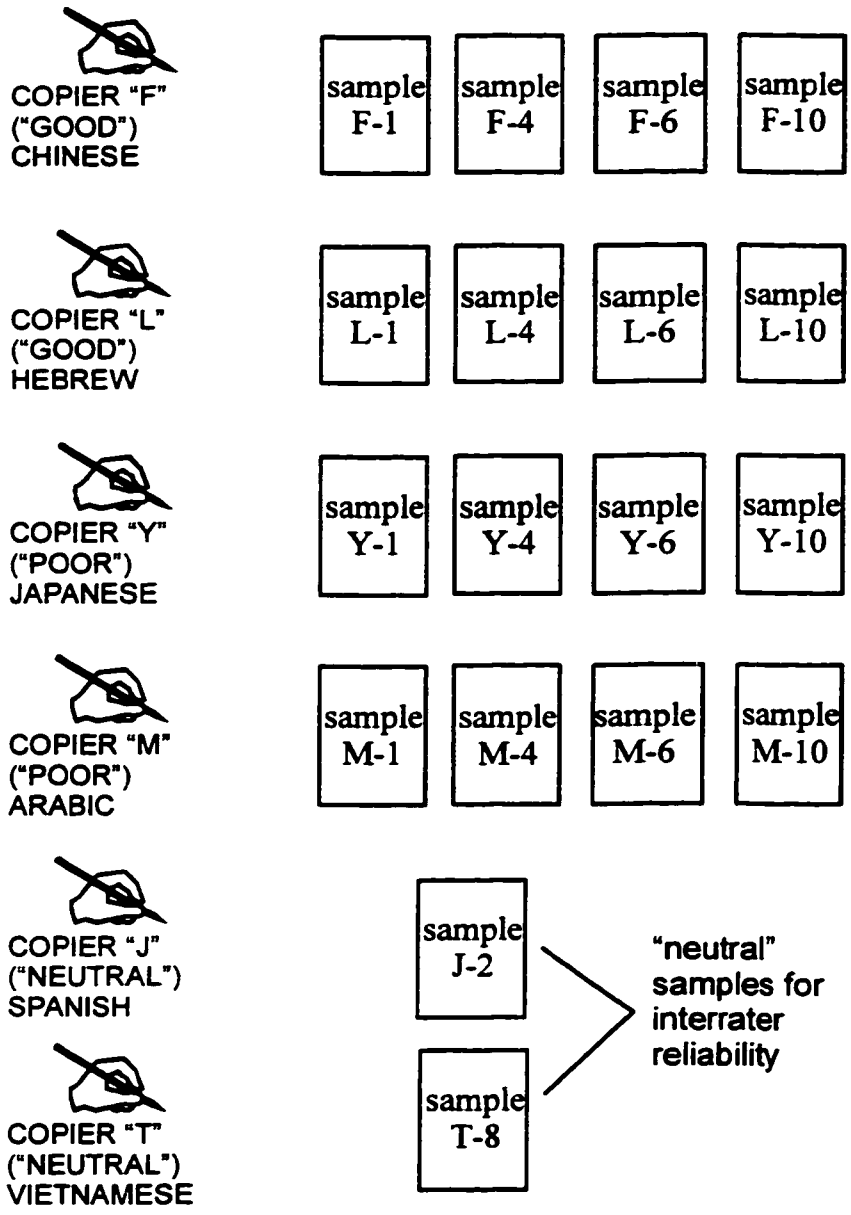

Figure 3.2. Copying the texts.

Another copier had particularly small and compact handwriting, which made the samples appear to be shorter than the other five. By re-copying with a wider margin while trying to write larger, the copier created samples that superficially appeared to be the same length as the others.

Figure 3.3 shows excerpts from the final samples. It should be noted that the following generalizations about "typical" handwriting from certain L1 learners are based on the experience of the researcher, not on any real analysis of cultural or national handwriting styles.

Copier F, whose L1 is Taiwanese, had a printed handwriting style that is "typical" of many learners from Taiwan, Korea, and Japan. It is characterized by small, neat printed letters that are remarkably uniform among individuals. This written style may derive from the primary educational systems in these countries, where the Roman alphabet is taught in elementary school as a system of romanization for Chinese, Japanese and Korean.

Copier L, whose L1 was Hebrew, had styles that might seem "normal" in a North American or European context.

Copier T was Vietnamese, and the handwriting style is somewhat cursive, and again is "typical" of many Vietnamese learners.

Copier J, who was from Venezuela, had a style that resembled typical North American handwriting.

Copier Y was from Japan, but has lived in Canada for over five years. The handwriting is not as neat and orderly as the writing of Copier F, but it retains some of the characteristics of "East Asian" handwriting.

Copier M was from Saudi Arabia, and has a "typical" Arab student's handwriting, which is characterized by inconsistent capitalization, inappropriate spacing between the letters and words, failure to stay on the line, and many ill-formed letters.

F **Chinese**	Pat became the undercover agent for the police. For them, she went to the drug dealer's house, but she was wearing a wire to get informations for the police.
L **Hebrew**	Pat has two daughters, addicted to the drugs. Her daugther take their grandchildren the drug dealer's farm every day.
T **Vietnamese**	The police met Pat in a Parking Lot. because of safty. The police wanted to look for this drug dealer since 6 years. They asked Pat to go to the
J **Spanish**	She met police secretely in a parking lot. The police were searching for this drug dealer for a long time. Pat wore a wire and went to the drug dealer's farm to look for her daughter
Y **Japanese**	Pat became the undercover agent for the police. For them, she went to the drug dealer's house, But she was wearing a wire to get informations for the police.
M **Arabic**	Pat has two daughters, addicted to the drugs. Her daughter take their grandchildren the drug dealer's farm every day.

Figure 3.3. Handwriting from the samples in rank order from best to worst.

3.3.2 Creating the Questionnaire

To ascertain the Stage 4 participants' range and depth of experience, as well as their views of writing assessment, a detailed four-page questionnaire was created (see Step 2 Questionnaire, Appendix D-10). In addition to personal information, the questionnaire elicited details of the length and type of ESL teaching of the participants, in the intensive program and elsewhere. One full page of the questionnaire asked how many times the participant had taught ESL-related courses at the institution, including both intensive and non-intensive courses.

The final section of the form dealt with the participants' views on the importance of certain aspects of ESL writing. Each item on the alphabetized list of 16 aspects was accompanied by a Likert-type scale from 1 to 6, with which the participants were to rate the "seriousness" of problems or errors with this aspect of writing for writers at the level they thought the samples to be. The purpose of this section was to see how the participants rated the importance of handwriting and how it related to the grades they gave samples in poor handwriting.

3.3.3 Creating the Packets

Experimental packets for the final steps of the study were prepared. Ten of each of four types of packets were made, with each type to be used by a different experimental group. Each packet contained the Step 1 Instruction sheet (See Appendix D-2) , a set of six writing samples, with grading slips attached, and the sealed Step 2 questionnaire. On the outside of the envelope, there were general instructions that asked the participants not to open the Step 2 questionnaire until Step 1 was completed. All these materials were put together in the packets to facilitate their distribution, and to protect anonymity. Since the

questionnaires and the grading slips had to be related by numerical code, it would have been very difficult to distribute the coded questionnaires without identifying the participants.

The Step 1 Instructions asked the participants to quickly look through the samples and determine what ESL level they were, and to write, on the Step 1 Instruction Sheet, the level and the degree of certainty they had about the level placement.

The texts that had been copied by hand are hereafter referred to as "samples". The samples were photocopied with blank spaces on the back of each for the Step 3 L1 Identification task. The samples were coded for identification purposes, and grading slips identical to those used in Stage 2 attached.

The arrangement of the samples in the packets is shown in Figure 3.4. For each sample, the letter represents the copier, and the number represents the text number. The samples were placed in coded envelopes in the order shown in the figure, from top to bottom. This ordering is by handwriting quality, not by text, so that all the groups get the same handwriting presented in the same order. In this way, the ordering effects found by Hughes and Keeling (1984) would apply equally to all packets.

The first two samples in each packet were the "neutral" handwriters, J and T, copying two middle-range texts. They were presented first because the neutral samples would be used for interrater reliability; it was therefore important that each participant judge them in the same way. The other four samples were the "experimental samples", and they occurred in a particular order in the packets. First was second-rated handwriter L, then sixth-rated M, then fifth rated Y, then first rated F. While the handwriting order was constant in all packets, the last four texts varied in position. The texts appeared in the

sequence loop 1→6→4→10, with a different starting position in each group. Appendices D-1 to D-10 show the contents of an experimental packet.

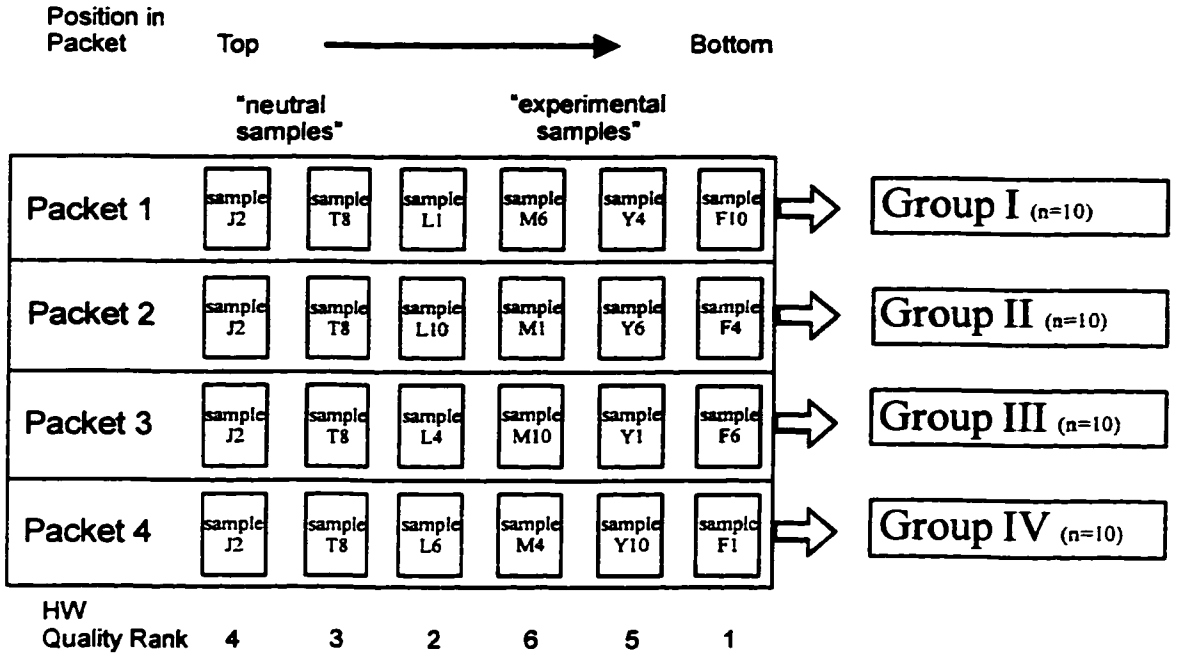

Figure 3.4. Ordering of the samples in each group's packet.

3.4 Stage 4: Main Experiment - Assessing the Handwritten Samples

3.4.1 Collecting The Data

Five data collection sessions were conducted, as it was impossible to schedule one which all 40 participants could attend. The sessions were held in both the morning and afternoon over the course of a week to accommodate the teaching schedules of the participants.

At each session, the participants were given the same verbal instructions, and an outline of the steps of the experiment was written on a blackboard at the front of the room. As each participant entered, he or she picked up an envelope containing one of the

four experimental packets. The pile of envelopes was arranged in a repeating sequence from Group 1 through to Group 4. In this way, the participants were randomly assigned to the four groups. The participants were told that the purpose of the experiment was to investigate how ESL teachers assessed writing holistically. They were asked not to discuss the experiment among themselves or with others until after all the data collection sessions were completed.

3.4.2 Step 1: Holistic Grading

For Step 1, the main group of participants were asked to remove the six writing samples from the envelope, and to read them all quickly. They were told that the samples were written in class, as a written reconstruction by students in the same class at the institution at which the main group were teachers. Next, they were asked to write the level they believed the writers were at in a space provided on the Step 1 Instruction sheet. The participants were told that it could be any level of any program (intensive, conversation, writing) at their institution. Having written down the level, the participants were asked to re-read the samples and circle a holistic grade for each on its attached grading slip. In the instructions, the participants were asked to grade the writing according to the criteria they would use to give formative feedback around the halfway point in the course.

After circling grades for all the samples, the participants were asked to tear off the grading slips and place them in a different envelope (placed on a desk in the centre of the room, within reach of all the participants), and put aside the samples. This step took between 15 and 20 minutes for most participants, with none finishing in less than 10 minutes. One participant took approximately 45 minutes to complete this step.

3.4.3 Step 2: The Questionnaire

As soon as each participant had placed his or her grading slips in the grading-slip envelope, he or she removed the sealed questionnaire (Appendix D-10) from the original experiment packet. Most of the participants took between 10 and 15 minutes to complete the questionnaire. Upon completion, the participants returned it to their envelopes and waited for instructions for Step 3.

3.4.4 Step 3: L1 Identification

As each participant completed Step 2, the researcher gave him or her a sheet of written instructions (Appendix D-11) for Step 3 which consisted of three questions to be answered for each of the six writing samples they had graded in Step 1. These instructions were not given orally, but were written so as not to influence the other participants who were still working on Step 2. First, the participants were asked to re-read the six writing samples, to make their "best guess" as to the L1 of the writer, and then to write that guess on the space provided on the back of each sample (Appendix D-12).

Second, the participants were asked to rate their level of certainty about their guesses on a four-point scale in which 1 was "very certain"; 2 was "somewhat certain"; 3 was "somewhat uncertain"; and 4 was "very uncertain". This scale was also printed on the back of each sample.

Third, the participants were asked to write down, on the spaces provided on the back of each sample, which features of the writing indicated the L1 they guessed. The participants were told that they could name a family or group of languages if they wished.

Step 3 took the longest for most participants and ranged from 20 to 30 minutes.

3.4.5 Completion of the Data Collection

As each participant finished Step 3, he or she was instructed to return all the samples and instructions to his or her original envelope and give it to the researcher. As each participant left the room, the researcher followed and thanked him or her in the hallway and repeated the request not to discuss any part of the experiment until all the data collection sessions had been completed. At this time, some participants made unsolicited comments about the various steps in the data collection.

3.4.6 Participant Debriefing

Several weeks after the data collection sessions, when the analysis of the data was complete, the researcher held two debriefing sessions at the institution to inform the participants of the results and the true purpose of the experiment. Following the presentation of the results, the researcher invited comments and questions, which were tape recorded. These sessions were attended by 24 of the original 40 participants. Another five participants, who couldn't attend the sessions, were informed on an individual basis and so gave their comments individually to the researcher.

Chapter 4: Results

4.1 Effects of Handwriting on Holistic Assessment

The first research question was, "What effect does handwriting quality and/or neatness have on the holistic assessment of L2 written work?" The data collected in the holistic grading in Stage 4 of the study were used to answer this question.

4.1.1 Interrater Reliability

To establish the degree of interrater reliability between the four main experimental groups in Stage 4, the scores assigned to the "neutral" samples, J2 and T8, were examined. The neutral samples appeared in all four experimental packets in the first and second position, respectively. Table 4.1 shows the mean group scores given to these samples and the standard deviations.

Table 4.1

Scores of Neutral Samples J2 and T8

	J2		T8	
	Mean	*SD*	Mean	*SD*
Group 1 (n = 10)	8.15	0.67	7.65	0.97
Group 2 (n = 10)	8.10	0.70	7.75	0.79
Group 3 (n =10)	7.70	0.95	7.45	0.60
Group 4 (n = 10)	7.90	0.66	7.70	0.75
All (n = 40)	7.96	0.75	7.64	0.77

An analysis of variance (ANOVA) was conducted on each set of scores to measure the degree of interrater reliability. The result for sample J2 is shown in Table 4.2, and the result for sample T8 is shown in Table 4.3.

Table 4.2

ANOVA of Scores for Neutral Samples J2 to Test Interrater Reliability

Variation	SS	*df*	MS	*F*
Between Groups	1.269	3	0.423	0.75
Within Groups	20.425	36	0.567	
Total	21.694	39		

Table 4.3

ANOVA of Scores for Neutral Samples T8 to Test Interrater Reliability

Variation	SS	*df*	MS	*F*
Between Groups	0.52	3	0.17	0.28
Within Groups	22.48	36	0.62	
Total	22.99	39		

There were no significant differences among the groups for either sample. Sample T8 had a p value of 0.84, and J2 had a p value of 0.53. This suggests that the groups were behaving in the same way when assessing these two samples. Because of the high level of interrater reliability found in these ANOVAs, it was assumed that the four experimental groups could be treated as one group, and further statistical tests could be conducted to answer the first research question. As a total of 10 subsequent comparisons were made on

the data provided by the main group of participants, the alpha was adjusted from .05 to .005, using the Bonferroni procedure.

4.1.2 Text Scores

Since the groups had been shown to have high levels of interrater reliability, the scores given to the other four "experimental" samples in each packet assessed by the main group were analysed for differences. The reader will recall that each of the four experimental texts was handwritten by a different copier in each experimental group. Table 4.4 shows the means and standard deviations of all the scores given to each text.

Table 4.4

Scores Given to Experimental Texts (N=40)

Text	Mean	*SD*
#4	7.31	0.89
#1	7.35	0.79
#10	7.45	0.84
#6	7.84	0.87

Table 4.5 shows the results of an ANOVA that was done on the experimental text scores, which revealed that there were no significant differences among the texts.

Table 4.5

ANOVA of Experimental Text Scores (N=40)

Variation	SS	*df*	MS	*F*
Between Groups	7.01	3	2.34	3.26
Within Groups	111.78	156	0.72	
Total	118.78	159		

4.1.3 Copier Scores

Having established that there were no significant differences in the experimental text scores, the same tests were done on all the scores given to the experimental samples that were handwritten by different copiers. Table 4.6 shows the means and standard deviations of all the scores given to each copier. Copier M, whose handwriting had been ranked the lowest by the selection group, received lower scores than the other copiers. Table 4.7 shows the results of an ANOVA done on the experimental copier scores. There were no significant differences among the copiers.

Table 4.6

Scores Given to Experimental Copiers (N=40)

Copier	HW Rank	Mean	*SD*
F	1	7.66	0.80
L	2	7.48	0.78
Y	5	7.61	0.98
M	6	7.20	0.84

Table 4.7

ANOVA of Experimental Copier Scores (N=40)

Variation	SS	*df*	MS	*F*
Between Groups	5.16	3	1.72	2.36
Within Groups	113.81	156	0.73	
Total	118.98	159		

4.1.4 Summary of the Effects of Handwriting on Holistic Assessment

Analyses done on the holistic scores given by the four groups of participants in Stage 4 showed a high degree of interrater reliability, based on the scores given to the neutral samples. No significant differences between the scores given to experimental texts or copiers were found, though the lowest ranked handwriter generally received lower scores.

4.2 Teaching Experience

The second research question was, "How does this effect relate to the length and range of teachers' experience?"

To determine if experience played a role in the holistic grading, the Step 2 Questionnaire data were used to regroup the participants for further analysis of the holistic assessment scores. This group of participants teach in a program in which the learners represent many different ethnic and linguistic backgrounds, and so it was expected that the length of experience with this student diversity might mitigate any effects of handwriting on holistic assessment. Therefore, the reported length of experience in multilingual adult ESL classrooms was considered the most relevant experience factor.

Table 4.8

Experience Groups

Experience	Definition	*n*
Experience Level 1	up to 4.5 years	10
Experience Level 2	5 to 7 years	10
Experience Level 3	8 to 10 years	10
Experience Level 4	more than 10 years	10

Based on the information given for how many years of experience each participant had in a multilingual adult ESL classroom, the participants were divided into

four blocks, from the least experienced to the most experienced. Each block contained 10 participants, as shown in Table 4.8. Table 4.9 shows the means and SDs produced by an analysis of the holistic scores assigned to the interrater reliability samples by the participants in each of the four experience groups. There was a tendency for the participants with less experience to give lower grades, but there were no significant differences between the four experience groups.

Table 4.9

Scores for Neutral Samples by Experience Level

	Level 1		Level 2		Level 3		Level 4	
	$n = 10$		$n = 10$		$n = 10$		$n = 10$	
Sample	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
J2	7.90	1.05	7.80	0.59	7.90	0.57	8.25	0.72
T8	7.40	0.81	7.45	0.90	7.80	0.54	7.90	0.77

The tendency for the participants at Experience Level 1 to give lower grades is clearer in Table 4.10 and Table 4.11, which shows the grades assigned to the other four copiers and texts. However, there were no significant differences between the experience groups in either the copier scores or the text scores.

Table 4.10

Scores Assigned to Experimental Copiers by Experience Level

	Level 1		Level 2		Level 3		Level 4	
Copier	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
F	7.50	0.78	7.50	1.03	7.75	0.42	7.90	0.91
L	7.35	0.82	7.60	0.99	7.60	0.66	7.35	0.67
Y	7.00	1.20	7.70	1.11	7.90	0.57	7.85	0.78
M	6.70	0.86	7.40	0.97	7.45	0.69	7.25	0.72

Table 4.11

Scores Assigned to Experimental Texts by Experience Level

	Level 1		Level 2		Level 3		Level 4	
Text	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
#1	7.10	1.02	7.40	0.99	7.45	0.44	7.45	0.60
#6	7.31	0.94	7.90	1.13	8.20	0.42	7.95	0.69
#4	7.10	0.97	7.45	0.90	7.40	0.74	7.30	1.03
#10	7.05	0.98	7.45	1.01	7.65	0.41	7.65	0.78

When all six samples were taken together, the tendency of the participants with the lowest level of experience to give lower grades was more pronounced, and appears to be applied to all samples, regardless of the handwriting. Table 4.12 shows the means and standard deviations of the all scores for all six samples, as well as for the four

experimental samples. As shown in Tables 4.13 and 4.14, ANOVAs revealed that there were no significant differences between the experience level groups in either case.

Table 4.12

Means and Standard Deviations of All Samples by Experience Level

	Level 1		Level 2		Level 3		Level 4	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
All 6 samples	7.31	0.97	7.58	0.92	7.73	0.58	7.75	0.81
4 experimental samples	7.14	0.95	7.55	0.99	7.68	0.59	7.59	0.80

Table 4.13

ANOVA of Scores Assigned to All Six Samples by Experience Group

Variation	SS	*df*	MS	*F*
Between Groups	7.49	3	2.50	3.61
Within Groups	163.29	236	0.69	
Total	170.78	239		

Table 4.14

ANOVA of the Four Experimental Samples Scores

Variation	SS	*df*	MS	*F*
Between Groups	6.79	3	2.26	3.15
Within Groups	111.99	156	0.72	
Total	118.79	159		

In summary, length of experience did not significantly affect graders' reactions to different qualities of handwriting, though it was found that the group of participants with the lowest level of experience tended to give lower grades to all samples than the two groups with the highest levels of experience.

4.3 Criteria For Grading

The third research question was, "How does this effect relate to teachers' views on writing assessment criteria?" The final page of the Step 2 Questionnaire provided data relevant to this research question. The main group of participants considered a list of errors or problems in writing that may occur at the level at which they thought the writers were. They then circled a number on a Likert-type scale from 1 to 6, corresponding to the level of "seriousness" of these errors or problems. 1 was "very serious" and 6 was "not serious."

4.3.1 Seriousness Ratings

Table 4.15 reflects the distribution of the seriousness ratings, showing the mean, the standard deviation, the median and mode for each of the criteria. On the questionnaire, the list was presented in alphabetical order, but in the tables, they are presented from most serious to least serious. Table 4.16 shows the actual distribution of the responses.

Problems with handwriting were rated as relatively non-serious by most of the main group participants, though the ratings were fairly evenly distributed across the scale. Other problems with mechanics, such as the spelling of common words, capitalization and punctuation were considered more serious by the participants.

Table 4.15

Distribution Statistics of Seriousness Ratings of Writing Assessment Criteria (N=40)

Criteria	Mean	*SD*	Median	Mode
Spelling: Common Words	2.20	1.02	2	3
Subject-Verb Agreement	2.33	1.12	2	2
Word Order	2.40	1.03	2	2
Capitalization	2.53	1.45	2	1
Organization	2.70	1.14	3	3
Pronouns	2.70	1.22	3	2
Run-On Sentences	2.75	1.08	3	3
Plurals	2.80	1.09	3	3
Punctuation	2.88	1.11	3	3
Fragment Sentences	2.88	1.16	3	2
Topic Sentences	3.25	1.48	3	2
Paragraph Cohesion	3.35	1.41	3	3
Handwriting	3.98	1.54	4	5
Prepositions	4.33	1.35	4	6
Articles	4.55	1.60	5	6
Spelling: Uncommon Words	4.83	1.01	5	5

Table 4.16

Distribution of Actual Responses for Seriousness Ratings

	Number of Responses					
Criteria	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5	Rating 6
Spelling: Common Words	12	12	13	2	1	
Subject-Verb Agreement	11	13	9	6	1	
Word Order	7	18	8	6	1	
Capitalization	12	10	10	3	3	2
Organization	7	9	16	5	3	
Pronouns	7	12	11	6	4	
Run-On Sentences	5	11	15	8		1
Plurals	4	12	16	4	4	
Punctuation	5	8	18	5	4	
Fragment Sentences	3	14	12	9	0	2
Topic Sentences	5	10	7	8	8	2
Paragraph Cohesion	3	9	12	6	7	3
Handwriting	2	7	6	8	9	8
Prepositions	1	2	9	9	9	10
Articles	2	4	5	4	9	16
Spelling: Uncommon Words		1	3	9	16	11

4.3.2 Experience Level and Seriousness Ratings

Non-parametric statistical tests found that there were significant differences in the seriousness ratings among teachers with different levels of experience.

Table 4.17 shows the means for all the seriousness ratings arranged by the four experience level groups that were used in the analysis in Section 4.2 (see Table 4.8). In general, the participants with higher levels of experience gave higher ratings, which means that they view all the writing assessment criteria as less serious than their less experienced colleagues.

Table 4.17

Means for All Seriousness Ratings by Experience Levels

	Experience Level 1 (n = 10)	Experience Level 2 (n = 10)	Experience Level 3 (n = 10)	Experience Level 4 (n = 10)
Mean Rank	2.93	2.90	3.21	3.58

A Kruskal-Wallis analysis of variance indicated significant differences among the groups ($\chi^2 = 38.36$, $df = 3$, $p < .0001$). Subsequent Mann-Whitney U tests on each pair of groups identified the significant differences, which are shown in Table 4.18. The participants with the highest level of experience (Experience 4) gave ratings that were significantly different than the groups with lower levels of experience.

Table 4.18

Mann-Whitney U Tests Between Pairs of Experience Groups

Pair	Experience 1	Experience 2	Experience 1	Experience 3
N	160	160	160	160
Mean Rank	2.93	2.90	2.93	3.21
U	12553.5		11476	
Z	-0.34		-1.84	
Pair	Experience 2	Experience 3	Experience 3	Experience 4
N	160	160	160	160
Mean Rank	2.90	3.21	3.21	3.58
U	11242.5		10910	
Z	-2.16		-2.49	
Pair	Experience 1	Experience 4	Experience 2	Experience 4
N	160	160	160	160
Mean Rank	2.93	3.58	2.90	3.58
U	9697.5*		9539*	
Z	-4.07		-4.26	

*$p < .0001$

Table 4.19

Means for Seriousness Ratings by Experience Level

Criteria	Experience Level 1	Experience Level 2	Experience Level 3	Experience Level 4
Spelling: Common Words	2.20	2.20	2.10	2.30
Subject-Verb Agreement	2.00	2.10	2.20	3.00
Word Order	1.80	2.40	2.30	3.10
Capitalization	2.60	2.50	2.30	2.70
Organization	2.10	2.60	3.20	2.90
Pronouns	2.80	2.40	2.30	3.30
Run-On Sentences	2.40	2.40	3.00	3.20
Plurals	2.80	2.50	2.80	3.10
Punctuation	3.10	2.80	2.80	2.80
Fragment Sentences	2.20	2.60	2.90	3.80
Topic Sentences	2.90	3.00	3.60	3.50
Paragraph Cohesion	2.80	2.80	3.80	4.00
Handwriting	4.30	3.80	4.20	3.60
Prepositions	4.10	3.70	4.20	5.30
Articles	4.50	3.80	4.30	5.60
Spelling: Uncommon Words	4.30	4.80	5.20	5.00

Table 4.19 shows the mean seriousness ratings for the individual writing assessment criteria given by four experience groups. No significant differences between the experience level groups were found for any of the individual writing assessment criteria.

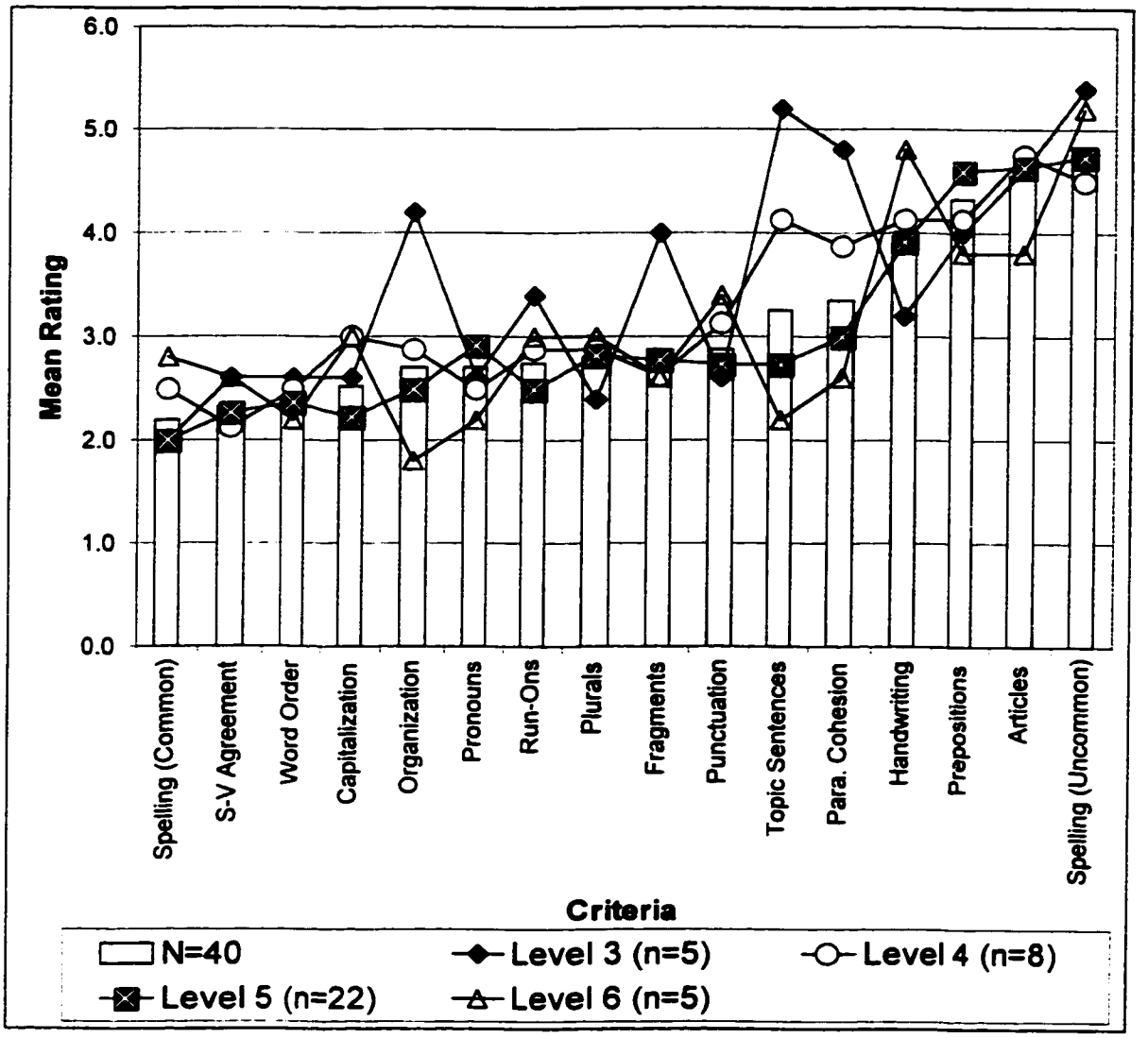

Figure 4.1. Seriousness Ratings by Guessed Level.

4.3.3 Guessed Level and Seriousness Ratings

Prior to indicating their seriousness rating for the writing assessment criteria, the participants were asked to recall the ESL level they had guessed the writer of the samples to be at. As can be seen in Table 4.20, there was little variation in the mean ratings among the different levels that were guessed, with the exception of three of the criteria: paragraph cohesion, organization, and topic sentences. These three are related to important teaching objectives that are introduced in Level 4 and developed through subsequent levels. Figure 4.1 shows the spread graphically.

4.3.4 Seriousness Ratings and Copier Scores

To find whether or not there was a relationship between the participants' views of the importance of penmanship and the grades given the "good" and "poor" copiers, the data were grouped according to the seriousness ranking of handwriting problems. Table 4.21 shows the means and SDs of the scores given by the participants who gave ratings of 1 to 6 for handwriting. The copiers are presented in the table in the order of the quality of their handwriting, as ranked by the selection participants in Stage 2.

As can be seen in Table 4.21, there are few apparent patterns or trends in the data that indicate a relationship between how serious the participants felt handwriting was as a writing assessment criterion and how they graded samples with better or worse quality handwriting. Moreover, because of the small number of participants in each cell of the table, no reliable tests could be done to test for significant differences.

Table 4.20

Seriousness Ratings by the Guessed ESL Level

Criteria	Level 3 (n = 5)	Level 4 (n = 8)	Level 5 (n = 22)	Level 6 (n = 5)
Spelling of Common Words	2.00	2.50	1.95	2.80
Subject-Verb Agreement	2.60	2.13	2.25	2.60
Word Order	2.60	2.50	2.38	2.20
Capitalization	2.60	3.00	2.17	3.00
Organization	4.20	2.88	2.57	1.80
Pronouns	2.60	2.50	2.86	2.20
Run-On Sentences	3.40	2.88	2.57	3.00
Plurals	2.40	2.88	2.60	2.50
Punctuation	2.60	3.13	2.76	3.40
Fragment Sentences	4.00	2.63	2.86	2.60
Topic Sentences	5.20	4.13	2.81	2.20
Paragraph Cohesion	4.80	3.88	3.05	2.60
Handwriting	3.20	4.13	3.85	4.80
Prepositions	4.00	4.13	4.67	3.80
Articles	4.60	4.75	4.62	3.80
Spelling of Uncommon Words	5.40	4.50	4.71	5.20

Table 4.21

Means and SDs of Scores by Handwriting Seriousness Rating

		Rating 1 (n=3)		Rating 2 (n=7)		Rating 3 (n=6)		Rating 4 (n=7)		Rating 5 (n=9)		Rating 6 (n=8)	
Copier	HW Rank	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
F	1	7.17	0.76	7.57	0.73	7.75	0.42	8.07	0.79	7.56	0.88	7.63	1.06
L	2	7.67	0.29	7.93	0.67	7.75	0.27	7.36	0.69	7.06	1.04	7.38	0.83
J	3	7.50	0.50	8.29	0.57	7.75	0.42	8.29	0.99	7.39	0.74	8.38	0.44
T	4	7.17	0.24	8.14	0.74	7.33	0.47	7.79	0.45	7.06	0.83	8.13	0.48
Y	5	7.67	0.29	7.86	0.69	7.67	0.61	8.07	0.73	7.00	1.27	7.63	1.27
M	6	7.50	0.50	7.57	0.61	7.25	0.61	7.43	0.67	6.83	1.20	6.94	0.86

4.3.5 Summary of Writing Assessment Criteria

The main group of participants differed in their seriousness ratings of the writing assessment criteria only according to their level of teaching experience. More experienced teachers rated all the criteria less seriously than their less experienced colleagues. The ratings for paragraph cohesion, organization, and topic sentences varied widely according to the level the teachers guessed the writers to be at. The other criteria were relatively consistent across the guessed levels.

4.4 L1 Identification

The final research questions was, "Can ESL teachers accurately guess the first languages of writers when handwriting is used as a cue?"

4.4.1 L1 Guesses

One of the final steps of the main experiment in Stage 4 was to make a guess as to the first language of the writer of each sample. These guesses were written by main group participants on blank spaces that were photocopied on the back of the six samples in the experimental packets (see Appendix D-12). In other spaces provided on the back of each sample, the participants recorded how certain they were of their guess, and wrote down the features of the sample of writing that led them to make their guess. This task was completed by 39 of the 40 main group participants. Each participant made guesses for six different samples. Table 4.22 shows the actual guesses made by the participants on the samples that were handwritten by each copier. The actual L1 of the copier is included in the table.

As can be seen in the table, many participants gave broad language groups, such as "Asian", "Oriental", or "European", as responses. Also, when a participant gave two or more L1s as a response, such as "Japanese or Korean" or "Italian or Spanish", it was recorded in one of the language groups.

Table 4.22

Guesses of Each Copier's L1

	F	L	T	J	Y	M	
Guess	Chinese	Hebrew	Vietnamese	Spanish	Japanese	Arabic	Total
Arabic		4		7	3	22	36
French	3	5	12	17	14	3	54
Spanish	3	11	10	8	10	4	46
Italian			1	1	1		3
Romance					1		1
German			1				1
Russian		1					1
European			1	1	1		3
Asian	7	5	2		2	2	18
Chinese	7	1	1		1	1	11
Japanese	12	6	6	1	4	4	33
Korean	3	1	2			2	8
Taiwanese	2					1	3
Vietnamese			1				1
Don't Know	2	5	2	4	2		15
Total	39	39	39	39	39	39	234

As many participants made multiple or language group guesses, all the responses were grouped for further analysis. Table 4.23 shows the percentage of participants who identified the copied texts in three broad groups: Arabic, Asian, and European. The participants were frequently correct in their guesses for the samples by Copiers F, J, and M, but were often wrong with Copiers L, T, and Y.

Table 4.23

Percentages of Participants Guessing L1 Groups by Copier

Copier	HW Rank	Actual L1	Don't Know	Arabic	Asian	European
F	1	Chinese	5.13%	0.00%	79.49%	15.38%
L	2	Hebrew	12.82%	10.26%	33.33%	43.59%
T	3	Vietnamese	5.13%	0.00%	30.77%	64.10%
J	4	Spanish	10.26%	17.95%	2.56%	69.23%
Y	5	Japanese	5.13%	7.69%	17.95%	69.23%
M	6	Arabic	0.00%	56.41%	25.64%	17.95%

4.4.2 Levels of Certainty

On the L1 identification task, the participants were asked indicate their level of certainty about their guesses, on a four point scale (1 = very certain; 2 = somewhat certain; 3 = somewhat uncertain; 4 = very uncertain). Overall, the participants were not very confident about their guesses: the mean level of certainty for all guesses was 2.58. The mean levels of certainty for the guesses made on the samples by each copier are

shown in Table 4.24. In the table, the data are presented in order of certainty, from most certain to least certain.

Table 4.24

Mean Ranks of Certainty for Guesses for Each Copier

Samples by Copier	Actual L1	HW Rank	Mean Rank of Certainty	*SD*
M	Arabic	6	2.26	0.99
F	Chinese	1	2.49	0.77
T	Vietnamese	4	2.51	0.87
L	Hebrew	2	2.59	0.89
J	Spanish	3	2.74	0.85
Y	Japanese	5	2.86	0.83

Participants were more confident when it came to texts copied by M, which had a mean certainty level of 2.26. On the other hand, texts copied by J and Y led to less certainty, with means of 2.74 and 2.86 respectively. There were no significant differences among these certainty levels of guesses for each copier.

However, as shown in Table 4.25, the participants who guessed that Copier M's samples were written by an Arabic speaker were very confident, when compared with those who thought the writer spoke a different L1. A Kruskal-Wallis analysis of variance indicated significant differences among the L1 guess groups ($\chi^2 = 13.81$, $df = 2$, $p <$

.001). Mann-Whitney *U* tests between the pairs of groups identified those differences, which are shown in Table 4.26.

Table 4.25

Certainty Levels for Guesses of Copier M's L1

L1 Guess	*N*	Mean Level of Certainty	*SD*
Arabic	22	1.82	0.85
European	7	2.71	0.76
Asian	10	2.88	0.93

Table 4.26

Mann-Whitney U tests of Certainty Levels for Guesses of Copier M's L1

	Arabic	Asian	Arabic	European	Asian	European
N	22	10	22	7	10	7
Mean Rank	1.82	2.88	1.82	2.71	2.88	2.71
U	44*		33.5		27.5	
Z	-2.71		-2.71		-0.76	

*$p < .01$

Note: The number of guesses for "European" was fewer than 8, so a normal distribution cannot be assumed.

Kruskall-Wallis tests were conducted on the certainty levels of guesses for the other five copiers, but no significant differences were found.

When the certainty levels for all the guesses of a particular L1 were analysed, some differences emerged. Table 4.27 shows the total number of guesses for each L1 group, and the mean level of certainty that accompanied those guesses. The participants who guessed that a sample was written by an Arabic speaker were more certain of those guesses than those who guessed other L1 groups. A Kruskal-Wallis test found that there were no significant differences among the L1 guess groups.

Table 4.27

Certainty Levels for All L1 Group Guesses

	N	Mean Rank	*SD*
Arabic	36	2.28	1.06
Asian	73	2.54	0.88
European	108	2.72	0.82

4.4.3 Handwriting Quality Rankings and L1 Guesses

The handwriting quality rankings of the selection group in Stage 2 provided another way to analyze the L1 guesses. Table 4.28 shows the mean handwriting ranking of the copiers to whom the L1 groups were attributed. The reader will recall that this ranking is from a high of 1 and a low of 6. Table 4.29 shows actual distributions of the rankings. There is a clear trend in the distribution that indicates that the participants are associating the samples with the poorer ranks of handwriting with Arabic writers, those with better handwriting with Asian writers.

Table 4.28

Handwriting Quality Rankings and L1 Guesses

L1 Group Guess	Mean of HW Quality Rank	*SD*
Asian	2.59	1.83
Don't Know	2.93	1.33
European	3.67	1.33
Arabic	5.08	1.36

Table 4.29

Distribution of Handwriting Quality Rankings and L1 Guesses

	HW Quality Ranking					
L1 Guess	1	2	3	4	5	6
Asian	31	13	12	1	7	10
European	6	17	25	27	27	7
Don't Know	2	5	2	4	2	
Arabic		4		7	3	22

4.4.4 Justifications for L1 Guesses

For the main group of participants, the final step in the L1 identification task was to write which features of the writing led the reader to guess that particular L1. All the responses were categorized and tallied, and they are presented in three tables: Table 4.30, by the original text of the samples; Table 4.31, broken down by the L1 Group guesses; and Table 4.32, by the handwriter who copied the sample. The tables do not reflect the fact that many participants wrote no justification for their choices at all. Most of the responses were in the form of short lists of one or two features, with little explanation.

The most common type of response, made 62 times for the 234 samples, was to quote an example from the text, with no further explanation. Two examples of this are, "she was worry" and "for/since". Presumably, these examples were intended to point out typical errors made by native speakers of the L1 in question.

The second largest category, with 57 occurrences, is "Verb Usage". The category is vague, but it covers a range of comments such as "trouble with verbs", "typical use of verbs", "tense problems", or "misuse of copula". It can be seen in Table 4.30 how evenly distributed the top two justifications are across all six copiers and the six texts. Most of the other responses were also evenly distributed among the texts although texts 2 and 8, which were the neutral samples used to establish interrater reliability and appeared in all the experimental packets, occasionally had different distributions than the four experimental texts.

Table 4.30

L1 Guess Justification by Text

	Number of Comments	Text 1	2	4	6	8	10
Gave example only	62	9	12	9	7	14	11
Verb Errors	57	8	8	10	9	9	13
Handwriting	55	13	3	13	12	6	8
Article Errors	30	3	3	4	9	7	4
Preposition Errors	30	5	6	5	4	5	5
Spelling	24	7	3	9	3		2
Sentence Structure	23	4	5	4	2	6	2
Plurals Errors	23	2	5	3	7	6	
Sentence Length	19	5		1	1	9	3
Word Order	19	3	2	3	1	4	6
Feel	15	1	3	3	2	2	4
Translating from L1	11	1	2	3	1	4	
Run-On Sentences	10	6		1	1		2

Table 4.30 (cont'd)

L1 Guess Justification by Text

	Number of Comments	Text 1	2	4	6	8	10
Punctuation	9		2	1	1	3	2
Fragment Sentences	7	1			1	2	3
Flow	7	1	2		2	1	1
Good	7		2		1	3	1
Organization	7	2	1		1	1	2
Sentence Boundaries	6	2	3			1	
Typical of L1	6	1	3			1	1
Sophisticated	5		1	1	2	1	
Capitalization	4		1	1		1	1
Choppy Sentences	3	3					
Not L1x, so must be L1y	2		1	1			
Indentation	1				1		

Table 4.31

L1 Guess Justifications By Guess

	Number of Comments	L1 Guess			
		Arabic	Asian	French	Spanish
Gave example only	62	6	11	27	14
Verb Errors	57	7	19	16	12
Handwriting	55	25	29	1	
Article Errors	30		13	7	9
Preposition Errors	30	1	15	9	5
Spelling	24	10	6	3	4
Sentence Structure	23	2	12	2	6
Plurals Errors	23	1	3	9	10
Sentence Length	19	2	13	1	3
Word Order	19	3	8	4	3
Feel	15	2	6	2	3
Translating from L1	11			9	2
Run-On Sentences	10	3			7

Table 4.31 (cont'd)

L1 Guess Justifications By Guess

	Number of Comments	L1 Guess			
		Arabic	Asian	French	Spanish
Punctuation	9	3	2	1	3
Fragment Sentences	7	2	4		1
Flow	7	2	2	2	1
Good	7		4	1	1
Organization	7	1	2	3	1
Sentence Boundaries	6	3	1	1	1
Typical of L1	6		2	4	
Sophisticated	5		3	2	
Capitalization	4	3			1
Choppy Sentences	3	1	2		
Not L1x, so must be L1y	2			1	
Indentation	1	1			

Table 4.32

L1 Guess Justification by Copier

	Number of Comments	Copier F	J	L	M	T	Y
Gave example only	62	6	12	11	8	14	11
Verb Errors	57	9	8	8	9	9	14
Handwriting	55	16	3	6	22	6	2
Article Errors	30	7	3	5	3	7	5
Preposition Errors	30	9	6	6	1	5	3
Spelling	24	3	3	7	11		
Sentence Structure	23	5	5	4	2	6	1
Plurals Errors	23	2	5	6	1	6	3
Sentence Length	19	4		1	4	9	1
Word Order	19	5	2	2	3	4	3
Feel	15	2	3	2	3	2	3
Translating from L1	11		2	1	2	4	2
Run-On Sentences	10	1		3	3		3

Table 4.32 (cont'd)

L1 Guess Justification by Copier

	Number of Comments	Copier F	J	L	M	T	Y
Punctuation	9	1	2	1	1	3	1
Fragment Sentences	7	1		2	1	2	1
Flow	7		2	2	1	1	1
Good	7	1	2	1		3	
Organization	7		1	1		1	4
Sentence Boundaries	6	1	3		1	1	
Typical of L1	6		3	1	1	1	
Sophisticated	5	2	1			1	1
Capitalization	4		1		2	1	
Choppy Sentences	3			1	2		
Not L1x, so must be L1y	2		1	1			
Indentation	1				1		

Most of the responses related to European languages were specific to one language, while those for Asian languages often were associated with the language group. As a result, Table 4.31 does not include the responses for European languages other than French and Spanish, so the total number of responses does not equal the sum of the L1 guesses column.

The third largest category, "handwriting", was written 55 times, and almost always for the texts that were guessed to be Asian or Arabic. Handwriting was cited on 22 of the texts copied by M, the Arabic writer, and 16 of those copied by F, the Chinese writer. Handwriting is also evenly distributed as a justification given in the four experimental texts that were copied.

Article and Preposition errors were frequently attributed to Asian writers, while spelling and capitalization were indicators mainly for Arabic writers. In all, 42 mechanics citations (handwriting, spelling, capitalization, punctuation, and indentations) were attributed to Arab writers, with 39 attributed to copier M's samples.

Sentence length, citations, including sentence boundaries, run-on sentences, fragment sentences, and "choppy" sentences were frequently used to justify L1 guesses. In general, the participants attributed short, choppy, or fragmented sentences to Asian writers, and long, run-on sentences to Spanish and occasionally to Arabic writers.

There were numerous vague responses such as "It feels like this L1" or "the flow", or "it is sophisticated", or "this is typical of this L1". These generally had no further explanation.

4.5 Unsolicited Main Group Participant Comments During Stage 4

Some participants made unsolicited comments after the data collection sessions. Regarding Step 1, the holistic grading, several participants expressed the view that although they understood the general context of the imagined writing task, it was difficult to grade the samples without a specific purpose in mind. In other words, these participants wanted to know what features of writing the teacher had been working on with the students prior to the task. Others commented that they are accustomed to responding to the individual development of the writer over the weeks of the course, suggesting that the developmental context of a particular writer is important information with which a teacher makes an assessment.

Commenting on Step 3, the L1 identification task, many participants remarked that it was very difficult to make guesses. Several mentioned handwriting as a deciding factor in their choices, citing the lack of other "typical" markers to help them. One said that it was hard not to be "biased" by the handwriting, reflecting a reluctance to rely on handwriting alone to identify the L1.

4.6 Main Group Participant Debriefing

During the tape-recorded debriefing sessions, most participants had comments and remarks to make on their performance and reaction to the various steps in the experiment.

Commenting on the preliminary task, in which the participants were asked to identify or guess the intensive ESL level of the writers, several participants reported being very self-conscious and concerned about their success on this question. They

believed, despite oral and written instructions to the contrary, that one of the purposes of the experiment was to measure the participants' ability to identify the level, and consequently they were very distracted by the question. Because many of the participants were aware that the content of the writing was a reconstruction of a video listening activity, they tried to recall the level at which the video is normally used. Those who had taught that level recently remembered the video and made their level decision quickly and easily.

Concerning Step 1, the holistic grading, there was a consensus among the participants that the level of language ability of the writers was very similar. Several participants said they were looking for differences among the writers but found only a limited range of writing ability, which seemed unusual in a "normal" class. Some remarked that all the papers had all the appropriate topic sentences, organizational patterns, and paragraph development that are major teaching points at the level they imagined the writers to be. Since all the papers met the major writing assessment criteria, the teachers had to look deeper than usual to distinguish them, as they were reluctant to give all the papers the same grade.

On the Step 3 L1 identification task, most participants reported that it was "obvious" that the handwriting of M was Arabic, and that of F Asian, but many expressed reluctance to go only by the handwriting to identify the L1. When some looked for further evidence to back up their impression based on the handwriting, they found little, and occasionally changed their guesses.

During one of the debriefing sessions, two participants reported that they were suspicious of the experiment's true purpose. Because they could not find evidence in the

samples that corroborated the first impression of the handwriting, they thought something had been manipulated in the samples, and that handwriting was involved. One of these participants explained that one of her mental strategies during the Step 1 holistic scoring was to try to imagine the writer. The handwriting was the most apparent clue for the text copied by M, but on more careful reading, she found herself thinking, "His spelling is very good for an Arabic speaker." She then started to notice mismatches between the errors and the handwriting in the other samples.

Other participants at that debriefing session said they did not notice a suspicious mismatch between the errors and the handwriting, and most expressed surprise upon learning of the manipulation of the samples.

Several participants suggested that they were guided in their guesses by mentally comparing the samples to work by the students in their current or recent classes.

Discussing handwriting and ESL writers, the participants expressed a common sentiment to the effect that experienced ESL teachers in multilingual classrooms have a high level of tolerance for poor handwriting and that they are able to read "just about anything". One participant remarked that ESL teachers should perhaps not be as tolerant because of the problems learners will encounter outside of the ESL classroom with poor handwriting. Another pointed out that poor handwriting may not be a significant factor in short, one-page written reconstructions done in a short period of time in class. On the other hand, in longer compositions, a grader's tolerance may be reduced because of irritation and fatigue.

Chapter 5: Discussion

5.1 Effects of Handwriting on Holistic Assessment

The first research question was, "What effect does handwriting quality and/or neatness have on the holistic assessment of L2 written work?" The results of this study suggest that there is no effect with this group of participants. As there were some differences between the scores assigned to the texts, regardless of the handwriting, it can be assumed that writing features other than penmanship were responsible for most of the variation among the groups, and that the amount of variation attributable to handwriting is outweighed by those other features.

The holistic scores did not correspond to the handwriting quality rankings of the samples. The samples produced by Copier M, whose handwriting was ranked the lowest in sixth place, generally received lower scores than the other samples, but the differences were not significant. The samples written by Copier Y, whose handwriting was ranked fifth, received scores that were comparable to those of Copier F, whose handwriting was ranked first in quality. Issues related to the design of the study might explain why no differences were found.

5.1.1 Handwriting Rankings

One problem in the design of the study was that the scale used to rank handwriting quality was not on an interval scale. The task given to the selection group of participants was to rank the 10 samples of handwriting. They were not trained or experienced in handwriting assessment and were asked only to rank the "quality" of the

samples, not to assign grades. The reader will recall that the mean rankings of the selection group were used to select the copiers. With such rankings, it is impossible to determine the real degree of difference of legibility between any two samples. In other words, the difference in legibility between the copiers, from first to sixth, cannot be assumed to be equal. This fact limited the statistical analysis that could be carried out, and makes it difficult to interpret the holistic grading in terms of handwriting "quality".

5.1.2 Range of Handwriting Quality

A possible explanation for why no differences were found is that there might not have been great enough differences in the quality of the handwriting. In other words, the good and poor handwriting was not good or poor enough to make a difference. All the samples, including those of Copier Y and M, were legible, and relatively short. As one of the main group of participants remarked during a debriefing session, irritation generated by poor or illegible handwriting is increased in longer compositions. In this study, short samples were used in order to keep the time required for data collection sessions at a reasonable length and to mimic typical in-class written reconstructions. In addition, in Stage 2, when the copiers were selected from the 10 handwriters, the top two and bottom two handwriting samples were rejected on the assumption that extreme differences in legibility might render the true purpose of the experiment obvious to the main group of participants.

5.1.3 Experienced ESL Teachers

The results of this study suggest that with the main group of participants, who are ESL teachers experienced in adult multilingual classrooms, handwriting quality makes no difference to assessment. In an L1 setting, there are established cultural norms for handwriting quality. An educator who teaches native speakers in North America probably reacts negatively or positively to handwriting when its quality or legibility falls outside the boundaries of what he or she considers "normal" handwriting. If such an educator encountered a text written by Copier M, he or she might think the writer had poor or childlike handwriting. On the other hand, an ESL teacher who is experienced in multilingual adult classrooms probably has different conceptions of what is "normal" or "acceptable" handwriting for L2 learners from different L1 backgrounds. The ESL teacher might not judge or even notice Copier M's handwriting in the same way as the L1 educator. In the debriefing sessions, some participants remarked that they believed that, as second language teachers, they were more tolerant of handwriting variation than teachers in L1 education contexts.

However, the conclusion that handwriting has no effect on holistic assessment cannot be generalized to different ESL teaching environments where the learners have a common L1, because the cultural norms mentioned above might be applied in those environments. Also, teachers who have limited exposure to the handwriting of adult international students might respond differently.

5.1.4 Fabricated Texts

In the early stages of designing the experiment, in consideration of the design features of L1 handwriting studies, the decision was made to use fabricated texts. Care was taken to remove errors in the texts that were "typical" of certain L1s, and make the texts as similar as possible. The comments of some of the main group of participants indicate that perhaps too much care was taken. Some said that they found the samples too close to each other in writing quality to be a plausible class set, and consequently became suspicious of the samples' authenticity, or looked more closely at the samples than they normally would to find differences. Others reported that in trying to visualize the anonymous writers, they looked for but failed to find textual features that corresponded to the handwriting. These remarks, combined with the fact that these tasks were done in a experimental setting, indicate that a number of the participants did not do the holistic grading task as they normally would. Using carefully selected or slightly modified authentic samples might have reduced some of these effects.

5.2 Experience and Handwriting

The second research question was, "How does this effect relate to the length and range of teachers' experience?"

It appears from the data that among the main group of participants in this study, length of experience has no bearing on the holistic scoring of different samples with different handwriting quality. It was found that the teachers in the group with the lowest level of experience (up to 4.5 years) gave lower scores for all the samples than the teachers in the two highest levels of experience (over 8 years), but this was not related to

handwriting quality alone. It might indicate a general tolerance for variation in writing skills that develops as the level of experience increases.

5.3 Writing Assessment Criteria

The third research question was, "How does this effect relate to teachers' views on writing assessment criteria?" It is clear from the data collected in this study that there is no relation between what the participants said about the seriousness of handwriting as an assessment criterion and the grades assigned to samples with a range of handwriting quality. None of the seriousness ratings for the other writing assessment criteria were related to the holistic scores, either.

An interesting pattern emerged in the grading criteria that reflects the nature of the language school that the main group of participants work at. Many of the seriousness ratings for the mechanics and grammatical criteria were widely distributed and were not related to the guessed level of the writing. However, three of the criteria, topic sentences, organization, and paragraph cohesion spread out corresponding to the level that the participants had guessed, while most of the others were clustered together, as shown in Figure 4.1. The spread probably reflects some of the primary teaching goals in writing in the intensive course at the school. At Level 3, there is very little attention paid to these three writing features, but as the level gets higher, so do the expectations of learner mastery of them. The other criteria are rarely mentioned explicitly in the curriculum of the school.

As with the holistic scoring discussed in Section 5.2, experience played a role in the distribution of all the criteria ratings. Carney (1973) found that less experienced

graders focussed more on mechanics, but in this study no differences were found in the seriousness ratings of mechanics criteria between the experience levels, though the teachers with lower levels of experience gave all the writing assessment criteria more serious ratings than those with the highest level of experience. Taking the experience findings from the holistic scoring and writing assessment criteria together suggests that in this study, the views of writing assessment of these teachers evolve and develop over time. The more experienced teachers generally gave higher holistic scores and less serious ratings than their less-experienced colleagues.

5.4 L1 Identification

The fourth research question was, "Can ESL teachers accurately guess the first languages of writers when handwriting is used as a cue?" In this study, it appears that they can in some cases. Most participants reported in the debriefing sessions that the L1 Identification task was very difficult. The way in which the participants responded to the task indicates that the guessing process is complex, but that handwriting is a very powerful cue in that process.

5.4.1 Guesses Typical of Linguistic Makeup of the School

The main group of participants identified the L1s in blocks or groups of languages that reflect the international make-up their school. An informal survey of the students conducted for recruiting purposes at the same time this study was carried out found that the L1s of the students are concentrated among only six languages, which are shown in Table 5.1. Fourteen other languages were reported, but none of them represented more than 2% of the population.

Table 5.1

First Languages of Students in the Intensive ESL Program

First Language	%
Spanish	21.20%
Japanese	17.51%
Chinese	16.59%
Arabic	12.90%
French	10.60%
Korean	6.91%

Since the participants were currently most familiar with students who spoke these L1s, it is not surprising that these were the most common guesses. The L1s for which there were only one or two guesses, such as Italian, Russian or Vietnamese, suggest that the participants may have had a speaker of these languages in their classes at that time or quite recently, and something in that particular text reminded them of that student. Some of the selection group participants in Stage 2 suggested that their L1 guesses for the word-processed texts were influenced by their recent experience with individual students.

5.4.2 L1 Group Guesses

The participants correctly identified Copier F as Asian 79.49% of the time, Copier J as European 69.23% of the time, and Copier M as Arabic 56.41% of the time. During

the debriefing sessions, many participants agreed that Copier F's handwriting, (see Figure 3.3) which is a neatly printed script, resembles the handwriting of many East Asian learners. They also agreed that Copier M's is typical of Arabic-speaking students. The participants made no comments on Copier J's handwriting, but the large number of European guesses for J and the other copiers indicates that when the handwriting is not identified as Arabic or Asian, it is identified as European, or more accurately "Other". The handwriting of J, T, Y, and L would not be atypical of native English speakers from North America. The wide variety of the guesses for T, Y, and L indicates that features other than the handwriting were used as cues for guessing, while the handwriting alone might have been a strong enough cue in the case of Copiers F and M.

5.4.3 Levels of Certainty

Many of the participants reported that the L1 identification task was very difficult and the relatively low levels of certainty reflect this. The fact that the levels of certainty for texts copied by M and F, and for all samples guessed to be Arabic or Asian were higher than those of the other copiers, supports the idea that typical Arabic and Asian handwriting is a powerful cue (See Table 4.25). The most compelling support comes from the large gap in certainty levels for the guesses made for Copier M's samples, where those who guessed Arabic had a mean certainty level of 1.82. compared to 2.71 for European guesses, and even less confidence for Asian guesses at 2.88 (see Table 4.26). This finding indicates that the strength of the handwriting as a cue eroded the certainty of the participants who found other features in the text to support a non-Arabic guess, or

conversely, those who guessed Arabic needed little extra evidence other than the handwriting.

5.4.4 Handwriting Quality Rankings

The comparison of guesses and handwriting quality rankings, shown in Table 4.28 and 4.29, suggests that the participants associate good, clear handwriting with Asian L1 speakers, and poor, illegible handwriting with Arabic speakers. The highest ranked handwriting was almost never guessed to be Arabic, and the lowest ranked was almost never thought to be Asian. This evidence suggests that the handwriting cue for L1 guesses is composed of two factors: national or regional style, and legibility, though it is impossible on the basis of this study to say which one has more influence.

5.4.5 L1 Guess Justifications

The final task completed by the main group of participants was to identify the features of samples that led them to make the L1 guesses they did. The nature and distribution of the justifications, shown in Tables 4.30, 4.31, and 4.32, provide more clues about the L1 guessing process.

Table 4.30 shows how evenly the responses are distributed among the texts. This confirms that the error types were indeed evenly distributed in the samples. The most common single justification was handwriting, though in the participant debriefing, many indicated that they were frustrated trying to find the typical markers that signaled certain L1s, and that they were forced to rely on handwriting more than they wanted to. The category "Gave example only" was used to classify the responses that consisted of an

example copied from the sample with no further explanation. "Verb errors" also covered a wide range of responses. Among the texts, handwriting is also evenly distributed, though the number of citations for Text 2 and 8 were somewhat lower than the other four. These texts were the neutral samples that appeared in the same form in each experimental packet and were used to establish interrater reliability, and they were most often guessed to be European.

In Table 4.31, the justifications are arranged by the L1 guess group. Handwriting was cited 25 times for Arabic guesses, and 29 times for Asian guesses, among the highest number of all the justifications.

Many of the other citations were often concentrated in one L1 group. After handwriting, the next most common justification were article errors, and these were most often associated with Asian writers, as were preposition errors. Spelling problems were mostly cited for Arabic speakers, while errors with plurals were equally frequent for French and Spanish writers.

A set of generalizations emerges from the distribution of the justifications, which is shown in Table 5.2. While this study provides no real evidence that these generalizations are true, the L1 guess justifications suggest that many of the teachers had a set of stereotypes about the writers who belong to these L1 groups.

Table 5.2

Generalizations about L1 Groups

L1 Group	Generalizations
Asian	Neat, clear handwriting
	Short, choppy sentences
	Problems with articles
	Problems with prepositions
Arabic	Poor handwriting
	Poor spelling
French	Problems with plurals
Spanish	Run-on sentences
	Problems with plurals

Other than the fairly specific concentration of run-on sentence citations for Spanish, the justifications for the European guesses were widely distributed and often vague. This suggests that the participants had to search for evidence in the samples to back up their guesses.

5.4.6 How the L1 Guesses Were Made

The evidence provided by the L1 guess justifications allow speculation on the process used by many of the participants in undertaking the L1 guessing task. The first steps in making the guess were influenced by the handwriting. If the handwriting style or

quality corresponded to the generalized parameters of Asian or Arabic writers, some participants made their decision at that point, with relatively high levels of certainty. Asked to justify their guesses, a large proportion cited the handwriting. Some were not comfortable relying solely on the handwriting, and looked more deeply into the text to find support for their guesses, which they found, in the form of errors with articles and prepositions, or short sentences for the Asian guesses, and spelling errors for Arabic guesses. It made no difference that the same errors occurred in the other samples in the experimental packet; the handwriting was powerful enough to draw attention to the errors that fit the generalization about that L1 group. A striking example of this is Text #1, which had six citations for run-on sentences, but also eight others for short sentences.

If the handwriting did not look Asian or Arabic, it fell into the "other" category, and closer reading was required. Because of the absence of L1 specific textual errors, the French or Spanish guesses were backed up with citations of errors such as "problems with verbs" that could be used to justify almost any L1 guess. In the end, these guesses were more tentative, with lower certainty levels. It was samples in the "other" category, copied by J, T, L, and Y, that generated 13 of the 15 "I don't know" guesses.

Of course, not all participants proceeded in this way. Some reported that they tried to ignore the handwriting. Some reported that they noticed the mismatch between the handwriting and the number and type of errors in the samples, and they carefully read and compared the samples to make their guesses. Some participants finished the task quickly, while others spent much longer looking for clues. Most of them agreed that it was difficult and somewhat frustrating.

5.5 Suggestions for Further Study

This study found that handwriting quality had no effect on the holistic scoring of ESL writing by teachers who are experienced with adult learners who speak a variety of L1s. As mentioned in Section 5.1, the effects might have been reduced by the narrow range of handwriting quality. Any further studies into the effects of appearance bias on scoring should ensure that the quality or legibility of the handwriting is quantitatively measured, and that the differences in quality cover a wide enough range to permit any handwriting effects that may exist to emerge. Many of the L1 handwriting studies reviewed in Chapter 2 used handwriting rating scales from the 1950s and earlier to assess the handwriting objectively. However, these scales reflect the cultural and educational norms of their era, and may or may not be applicable to the handwriting styles of adult L2 learners from various national and cultural backgrounds.

No effects were found for this group of teachers, but that does not mean that effects would not be found in different ESL settings. The evidence from L1 handwriting studies suggests that the effect is widespread, and teachers who do not have experience in adult multilingual classrooms might react to poor handwriting in the same way L1 educators do. It would be interesting to compare these results to a similar study using teachers from a classroom environment where the L2 learners share the same L1.

It seems clear that handwriting influences L1 identification of anonymous writers, though the phenomenon of national or regional writing styles has barely been investigated at this time. Future studies could use large numbers of authentic samples from students from many L1 or regional backgrounds to confirm the existence of generalized writing styles of Asian and Arabic writers.

References

Arnold, V., Legas, J., Obler, S., Pacheco, M., Russell, C., & Umbdenstock, L. (1990). *Do students get higher scores on their word-processed papers?*. Whittier, CA: Rio Hondo College.

Ball, W. (1986) Writing English script: an overlooked skill. *ELT Journal, 40*, 291 – 298.

Brennan, E. M., & Brennan, J. S. (1981). Accent scaling and language attitudes: Reactions to Mexican American English speech. *Language and Speech, 24*, 207 – 221.

Briggs, D. (1970). Influence of handwriting on assessment. *Educational Research, 13*, 50 - 55.

Briggs, D. (1980). A study of the influence of handwriting upon grades using examination scripts. *Educational Review, 32*, 185 - 193.

Brosell, G. (1986). Current research and unanswered questions in writing assessment. In K. Greenberg, H. Wiener, & R. Donovan (Eds.) *Writing assessment: Issues and strategies* (pp. 168 - 182). New York: Longman.

Burt, M. K. & Kiparsky, C. (1972) *The gooficon*. Rowley, MA: Newbury House.

Carlson, S., & Bridgeman, B. (1986). Testing ESL student writers. In K. Greenberg, H. Wiener, & R. Donovan (Eds.) *Writing assessment: Issues and strategies* (pp. 126 - 152). New York: Longman.

Carney, H. (1973). An inquiry into criteria for composition evaluation in English as a second language. *Dissertation Abstracts International, 34* (8), 5139-A.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: a critical overview. *Research in the Teaching of English, 18*, 65 - 81.

Chase, C. (1968). The impact of some obvious variables on essay scores. *Journal of Educational Measurement, 5*, 315 - 318.

Chase, C. (1979). The impact of achievement expectations and handwriting quality on scoring essay tests. *Journal of Educational Measurement, 16*, 39 - 42.

Chase, C. (1986). Essay test scoring: interaction of relevant variables. *Journal of Educational Measurement, 23*, 33 - 41.

Cumming, A. (1990). Expertise in evaluation second language compositions. *Language Testing, 7*, 31 – 51.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*, 1 - 16.

Diederich, P. (1974). *Measuring growth in English.* Urbana IL: National Council of Teachers of English.

Eames, K. & Loewenthal, K. (1990). Effects of handwriting and examiner's expertise on assessment of essays. *Journal of Social Psychology, 130*, 831-833.

Emerling, F. (1991). Identifying ethnicity and gender from anonymous essays. *Community College Review, 19*, 29 -33.

Farquharson, M. (1988, March). *Ideas for teaching Arab students in a multicultural setting.* Paper presented at TESOL '88, Chicago, IL.

Freedman, S. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71*, 328 - 338.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*, 65 - 89.

Greenberg, K., Wiener, H., & Donovan, R. (Eds.) (1986). *Writing assessment: Issues and strategies*. New York: Longman.

Homburg, T. (1984). Holistic evaluation of ESL compositions: can it be validated objectively? *TESOL Quarterly, 18*, 87 - 107.

Huck, S. & Bounds, W. (1972). Essay grades: and interaction between graders' handwriting clarity and the neatness of examination papers. *American Educational Research Journal, 9*, 279 - 283.

Hughes, D., Keeling, B., & Tuck, B. (1983). The effects of instructions to scorers intended to reduce context effects in essay scoring. *Educational and Psychological Measurement. 43*, 1047 - 1050.

Hughes, D. & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement, 21*, 277 - 281.

Huot, B. (1990). Reliability, validity, and holistic scoring: what we know and what we need to know. *College Composition and Communication, 41*, 201 - 213.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: a practical approach*. Rowley, MA: Newbury House.

Klein, S. & Hart, F. (1968). Chance and systematic factors affecting essay grades. *Journal of Educational Measurement, 5*, 197 - 206.

Loewenthal, K. (1980). What does handwriting tell us about personality? *New Society, 51*, 544 – 546.

Markham, L. (1976). Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal, 13*, 277 - 283.

Marshall, J. (1972). Writing neatness, composition errors, and essay grades reexamined. *The Journal of Educational Research, 65*, 213 - 215.

Massey, A. (1983). The effects of handwriting and other incidental variables on GCE 'A' level marks in English literature. *Educational Review, 35*, 45 –50.

McColly, W. (1970). What does educational research say about the judging of writing ability? *Journal of Educational Research, 64*, 148 - 156.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning, 45*, 73 - 97.

Nevarez, H., Berk, V., & Hayes, C. (1979). The role of handwriting in TESOL. *TESOL Newsletter, 13*, 25 - 26.

Odlin, T. (1989). *Language transfer*. Cambridge: Cambridge University Press.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective writing tests to evaluate ESL writing ability. *TESOL Quarterly, 17*, 651 - 671.

Peterson, E. & Lou, W. (1991*). The Impact Of Length On Handwritten And Wordprocessed Papers*. Paper presented at the Annual Meeting of the Oregon Educational Research Association, Portland, OR.

Powers, D., Fowles, M., Farnhum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*, 220 - 233.

Santos, S. & Suleiman, M. (1993). Teaching English to Arabic-speaking students. *Proceedings of the National Association for Bilingual Education Conferences, Tucson, AZ, 1990; Washington, DC, 1991*, 175-180.

Sloan, C. & McGinnis, I. (1982). Effect of handwriting on teacher's grading of high school essays. *Journal of the Association for the Study of Perception, 17*, 15 - 21. ERIC Document # ED 220 836

Soloff, S. (1973). Effect of non-content factors on the grading of essays. *Graduate Research in Education and Related Disciplines, 6*, 44 - 54.

Sprouse, J. & Webb, J. (1994) The Pygmalion Effect and its Influence on the Grading and Gender Assignment on Spelling and Essay Assessments. ERIC Document # ED 374 096

Sweedler-Brown, C. (1992). The effect of training on the appearance bias of holistic essay graders. *Journal of Research and Development in Education, 26*, 24 -29.

Thompson-Panos, K. & Thomas-Ruzic, M. (1983). The least your should know about Arabic: Implications for the ESL writing instructor. *TESOL Quarterly, 17*, 609 - 623.

Varonis, E., & Gass, S. (1982). The comprehensibility of nonnative speech. *Studies in Second Language Acquisition, 4*, 114 - 136.

APPENDIX A: Consent Forms

CONSENT FORM TO PARTICIPATE IN RESEARCH -1

This is to state that I agree to participate in a program of research being conducted by Kevin Stanley as part of his M.A. thesis under the supervision of Dr. Joanna White of the TESL Centre at Concordia University.

A. PURPOSE

I have been informed that the purpose of the research is to study how ESL teachers assess the writing of ESL students.

B. PROCEDURES

In the research, participants will copy several samples of writing by hand. Completion of this task should take about one hour. The participants names will not appear on the papers, and therefore there will be no way to identify each writer. After the analysis of the data, the researcher will inform all participants of the results.

C. CONDITIONS OF PARTICIPATION

- I understand that I am free to withdraw my consent and discontinue my participation at any time without negative consequences.
- I understand that my participation in this study is CONFIDENTIAL.
- I understand that the data from this study may be published.
- I understand the purpose of this study and know that there is no hidden motive of which I have not been informed.

I HAVE CAREFULLY STUDIED THE ABOVE AND UNDERSTAND THIS AGREEMENT. I FREELY CONSENT AND AGREE TO PARTICIPATE IN THIS STUDY.

NAME (please print) _____

SIGNATURE _____

WITNESS SIGNATURE _____

DATE _____

CONSENT FORM TO PARTICIPATE IN RESEARCH -2

This is to state that I agree to participate in a program of research being conducted by Kevin Stanley as part of his M.A. thesis under the supervision of Dr. Joanna White of the TESL Centre at Concordia University.

A. PURPOSE

I have been informed that the purpose of the research is to study how ESL teachers assess the writing of ESL students.

B. PROCEDURES

In the research, participants will assess several samples of writing. Completion of this task should take about one hour. All assessment data will be treated in the strictest confidence, and confidentiality will be assured through the use of numerical coding on all the documents. In this way, the researcher will not know how individuals scored. Also, only group results will be presented or published. After the analysis of the data, the researcher will inform all participants of the results.

C. CONDITIONS OF PARTICIPATION

- I understand that I am free to withdraw my consent and discontinue my participation at any time without negative consequences.
- I understand that my participation in this study is CONFIDENTIAL.
- I understand that the data from this study may be published.
- I understand the purpose of this study and know that there is no hidden motive of which I have not been informed.

I HAVE CAREFULLY STUDIED THE ABOVE AND UNDERSTAND THIS AGREEMENT. I FREELY CONSENT AND AGREE TO PARTICIPATE IN THIS STUDY.

NAME (please print) _____

SIGNATURE _____

WITNESS SIGNATURE _____

DATE _____

CONSENT FORM TO PARTICIPATE IN RESEARCH-3

This is to state that I agree to participate in a program of research being conducted by Kevin Stanley as part of his M.A. thesis under the supervision of Dr. Joanna White of the TESL Centre at Concordia University.

A. PURPOSE

I have been informed that the purpose of the research is to study how ESL teachers assess the writing of ESL students.

B. PROCEDURES

In the research, participants will assess several samples of writing, and then complete a questionnaire. The questionnaire includes questions concerning personal information, teaching experience and writing assessment criteria. Completion of these tasks should take about one hour. All assessment and questionnaire data will be treated in the strictest confidence, and confidentiality will be assured through the use of numerical coding on all the documents. In this way, the researcher will not know how individuals scored. Also, only group results will be presented or published. After the analysis of the data, the researcher will inform all participants of the results.

C. CONDITIONS OF PARTICIPATION

- I understand that I am free to withdraw my consent and discontinue my participation at any time without negative consequences.
- I understand that my participation in this study is CONFIDENTIAL.
- I understand that the data from this study may be published.
- I understand the purpose of this study and know that there is no hidden motive of which I have not been informed.

I HAVE CAREFULLY STUDIED THE ABOVE AND UNDERSTAND THIS AGREEMENT. I FREELY CONSENT AND AGREE TO PARTICIPATE IN THIS STUDY.

NAME (please print) _____

SIGNATURE _____

WITNESS SIGNATURE _____

DATE _____

APPENDIX B: The Six Texts

Text #1

Pat Morrison, a 44 year old grandmother, who has two daughters addicted to drugs. Sometimes they took their children with them to drug dealer's house. She is decided call the police because she wanted her daughters to jail for a drug program.

She met police in secretly. Pat voluntered for be an under cover agent. She visited the drug dealer's house, she said she is looking for her daughters, she is wearing a wire.

Her daughters introduced her to Jimy, the drug dealer gang's leader. Pat told Jimy her friend wanted to buy 1 pounds of drug for $15,000. Jimy didn't give her right away it wasn't ready and he needs chemicals, so Pat got for him chemicals.

Jimy took Pat to the secret lab. When they were driving, Jimy's man asked Pat for the wire. Pat was very fraightened, but they were joking.

The police arrested Jimy and his gang. Jimy is going to jail for two years and now Pat's daughters are off drugs.

Text #2

Pat is 44 years old and a grandmother. Her 2 daughters are adicted to some drug. Pat worries about the grandchildren because they are sometimes dirty. She called to the police because she wanted her daughters in the prison so they can stop to take the drug.

She met police secretely in a parking lot. The police were searching for this drug dealer for a long time. Pat wore a wire and went to the drug dealer's farm, to look for her daughters.

She met Jimmy who is the drug dealer. She said, "My friend wants to buy one pounds of your drug for $15000." Jimmy said, "I need one bottle of chemical." Then Pat gave him the chemicals from the police.

After three weeks, they went to the lab in a secret place. Jimmy joked about the wire, and Pat was very scared. The next day, the police arrested all the gang. Jimy in prison for two long years and Pat is happy because her daughters quitted the drugs.

Text #4

Pat is forty four year old woman and grandmother. Her two daughter are drug addicted. Her daugthers take the grandchildren the drug dealer's farm. She is very worry and wants them to quit it.

In the parking lot, she was met by the police. The police wanted her to undercover infiltate Jimmy's gang. It is very dangerous, but Pat went Jimmy's house with police wire.

Her daughter introduce Jimmy to Pat. Pat said her friend wanted one pound of the drug for fifteen thousand dollars. Jimmy said he does not have it, because he does not have the special cemichal. So Pat asked the police for the cemichal and brought it to Jimmy.

In three weeks, Jimmy is trusted Pat. They went together to the lab. Jimmy was joking about the wire and Pat did not understand it was joking, so she was nervous. In the next day, Jimmy's gang was arrested by police, and they went to the jail. Pat's daughters are now free for the drugs.

Text #6

Pat Morrison was a mother and a grandmother, but her daughter were addicted to the drugs. Pat was worry about her grandchildren, so called the police because she wanted the daughters go to the prison for the drug program.

Pat became the undercover agent for the police. For them, she went to the drug dealer's house, but she was wearing a wire to get infornations for the police.

She introduced to the gang's leader, Jimmy. Pat said to Jimmy she wanted to buy one pound of the drugs for $15,000. Jimmy couldn't give her the drugs because he needed to the chemicals, so Pat got the chemicals for him.

Then Jimmy took Pat to the secret lab for manufatturing the drugs. Jimmy and the other drug dealer said to Pat they wanted the wire. Pat was so scared, but they were joking.

Then the police arrested Jimmy and the gang. He is going to the prison for the two years and now Pat's daughters are not taking the drugs.

Text #8

Pat Morisson is 44 and a grandmother. Her two daughters are using a drug. Sometimes, they take Pat's grandchildren to farm of the drug dealer, and the children aren't well. She wants to help them stop the drugs, so she called to the police.

The police met Pat in a parking lot, because of safty. The police wanted to look for this drug dealer since 6 years. They asked Pat to go to a drug dealer's farm with a wire.

Her daughters introduced Pat for the drug dealer. Pat was wanting to buy a lot of drug for $15,000. Jimmy said he doesn't sell her, because he doesn't have chemical. The police gave Pat the chemical for the drug dealer.

After three weeks, the drug dealer trusted to Pat. He took her to the lab to make drug. Pat was afraid because the drug dealer joked about the wire. The polices arested the gang and the drug dealer then went to the jail. Pat's daughters now are stopped the drugs.

Text #10

Pat has two daughters, addicted to the drugs. Her daugther take their grandchildren the drug dealer's farm every day.

Pat was very worried because the kids are not clean or eating. For her daughters to quit it, she called police.

Pat was meeting the police in the safety parking lot. The police wanted Pat infiltration of Jimmy's (drug dealer) gang. Pat decided to visit to Jimmy's house while she was wearing the police wire, even though dangerous.

Pat asked to Jimmy for $15,000 of drug for her friends. Jimmy haven't the drug in hand because he needs the chemical. The police listened and gave Pat chemical for giving to Jimmy.

Because of this, Jimmy trusted Pat. They drove in the truck to the drug lab. Pat was nervuous because Jimmy talked about the police wire. However he was joking, then. In the morning of dawn, Jimmy and the gang was arrested. Moreover, the daughters of Pat quitted the drugs.

APPENDIX C: Instructions to the Copiers

INSTRUCTIONS FOR COPYING THE STORIES

1. Copy the stories exactly. There are spelling and grammar mistakes in the story. Be careful not to correct the mistakes. Pay attention to commas(,), periods (.) and quotation marks (" "). After you give me your copies, I have to check them to make sure the mistakes are the same. Please use a dark pencil, so it is easy to fix them, and easy to photocopy.

2. Please double space when you copy. This means don't write on every line. Leave one empty line between each line of writing.

3. Your story must fit on one page. If you need more than one page, please try to write smaller. If your writing is too small, please try to write bigger.

4. Please use the paper I give you. The stories must be the same size.

5. Don't be too careful with your writing. I want the stories to look natural and normal. I want the stories to look like stories that students wrote in class, not at home. Don't try to write perfectly. Write quickly if you can.

6. Please don't talk about this secret experiment with anyone. It is important that the teachers don't know anything about it. It is possible that your friends or classmates could talk to their teachers about it by accident. So please be patient. I hope that the experiment will be finished before the end of the course. After that, it will be OK to talk about it.

Thank you very much for your help.

APPENDIX D-1: General Instructions

(Attached to the Outside of the Experimental Packet Envelopes)

GENERAL INSTRUCTIONS

There are three steps in this experiment. In Step 1, you will be asked to assess some student writing samples. In Step 2, you will be asked to complete a questionnaire. In Step 3, you will asked to assess the same writing samples for a different purpose.

It is important for this study that you complete the steps in order. Please don't open the questionnaire until you have completed Step 1, and I have collected the grading slips. I will distribute the instructions for Step 3 after you have completed Step 2 Questionnaire.

When you have finished, please return all materials and instructions to this envelope (except those that I collect from you).

This envelope contains:

- Instructions for Step 1
- 6 writing samples with grading slips attached
- The Step 2 Questionnaire

About Confidentiality

The materials in this package are alphanumerically coded. The researcher will be able to identify individuals only by that code. The purpose of the code is to allow cross-referencing between the writing assessment data and the questionnaires.

Only group results and data will be analyzed or presented. In other words, individual performance is not the focus of this study, and in no way are individual results going to be compared or rated.

Thank you for assistance with this study. You time and patience are appreciated.

APPENDIX D-2: Step 1 Instructions

Step 1

Instructions

The samples in this packet are written reconstructions of a listening task done at this school. Please assume that these samples were written by students in a class you were teaching, between week 3 and week 8.

1. Please look the samples over. What level do you think they came from?

_____

2. How certain are you that this writing comes from the level you have chosen?

_____ Very Certain
_____ Somewhat Certain
_____ Somewhat Uncertain
_____ Very Uncertain

3. On the slips of paper attached to each sample, please circle a holistic grade for the writing. Assume that the grading is for a formative evaluation (not a final exam) and that the writing was done at some point between weeks 3 and 8 of the course.

4. Tear off the slips when you have finished and set aside the samples which you will need in Step 3. Once I have collected your grading slips, you can proceed to Step 2.

Code: _____

APPENDIX D-3: Grading Slip

(Stapled to Each Sample)

Group ____ Code: _____

Please circle a grade:

0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0

5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0 9.5 10

APPENDIX D-4: Sample J2

(from Group 2 Experimental Packet)

Pat is 44 years old and a grandmother. Her 2 daughters are addicted to some drug. Pat worries about the granddildren because they are sometimes dirty. She called to the police because she wanted her daughters in the prison so they can stop to take the drug.

She met police secretely in a parking lot. The police were searching for this drug dealer for a long time. Pat wore a wire and went to the drug dealers room to look for her daughters. She met Jimmy who is the drug dealer. She said "My friend wants to buy one pounds of your drug for $ 15.000." Jimmy said, "I need one bottle of chemical." Then Pat gave him the chemicals from the police.

After three weeks, they went to the lab in a secret place. Jimmy joked about the wire, and pat was very scared. The next day, the police arrested all the gang. Jimmy in prison for two long years and Pat is happy because her daughters quitted the drugs.

APPENDIX D-5: Sample T8

(from Group 2 Experimental Packet)

Pat Morisson is 44 and a grand mother. Her two daughters are using a drug. Sometimes, They take Pat's grandchildren to farm of the drug dealer, and the children aren't well. She wanted to help them stop the drugs so she called to the police.

The police met Pat in a Parking lot. because of safty. The police wanted to look for this drug dealer since 6 years. They asked Pat to go to the drug dealer's farm with a wire. Her daughters introduced Pat for the drug dealer. Pat was wanting to buy a lot of drug for $15-000. Jimmy said he doesn't sell her, because he doesn't have Chemical. The police gave Pat the chemical for the drug dealer.

After three weeks, the drug dealer trusted to Pat. He took her to the lab to make drug. Pat was afraid because the drug dealer talked about the wire. The Polices arested the gang and the drug dealer then went to the Jail. Pat's daughters now are stopped the drugs.

10

APPENDIX D-6: Sample L10

(from Group 2 Experimental Packet)

Pat has two daughters, addicted to the drugs. Her daugther take their grandchildren the drug dealer's farm every day.

Pat was very worried because the kids are not clean or eating. For her daughters to quit it, she called police.

Pat was meeting the police in the safety parking lot. The police wanted Pat infiltration of Jimmy's (drug dealer) gang. Pat decided to visit to Jimmy's house while she was wearing the police wire, even though dangerous.

Pat asked to Jimmy for $15.000 of drug for her friends. Jimmy haven't the drug in hand because he needs the chemical. The police listened and gave Pat chemical for giving to Jimmy.

Because of this, Jimmy trusted Pat. They drove in the truck to the drug lab. Pat was nervuas because Jimmy talked about the police wire. However he was joking, then. In the morning of dawn, Jimmy and the gang was arrested. Moreover, the daughters of Pat quitted drugs.

APPENDIX D-7: Sample M1

(from Group 2 Experimental Packet)

Pat Morrison, a 44 year old grandmother, who has two daughters addicted to drugs. Sometimes they took their children with them to drug dealer's house. She is decided call the police because she wanted her daughters to jail for a drug program.

She met police in secretly Pat volunteered for be an under cover agent She visited the drug dealer's house. She said she is looking for her daughters, She is wearing a wire.

Her daughters introduced her to Jimy, the drug dealer gong's leader Pat told Jimy her friend wanted to buy 1 pounds of drug for $15,000 Jimy didn't give her right away it wasn't ready and he needs chemicals, so Pat got for him chemicals Jimy took Pat to the secret lab When they were driving Jimy's man asked Pat for the wire Pat was very fraightened, but they were joking

The police arrested Jimy and his gang Jimy is going to jail for two years and now Pat's daughters are off drugs.

APPENDIX D-8: Sample Y6

(from Group 2 Experimental Packet)

Pat Morrison was a widow and a grandmother, but her daughter were addicted to the drugs. Pat was worry about her grandchildren, so called the police because she wanted the daughters go to the prison for the drug program.

Pat Become the undercover agent for the police. For them, she went to the drug dealer's house, But she was wearing a wire to get infomations for the police.

She introduced to the gang's leader, Jimmy. Pat said to Jimmy she wanted to buy one pound of the drugs for $15,000. Jimmy couldn't give her the drugs because he needs to the chemicals so Pat got the chemicals for him.

Then Jimmy took Pat to the secret hall for manufacturing drugs. Jimmy and the other drug drug dealer said to Pat they wanted the wire. Pat was so scared, but they were police.

Then the police arrested Jimmy and the gang. He's going to the prison for the two years and now Pat's daughters are not taking the drugs.

APPENDIX D-9: Sample F4

(from Group 2 Experimental Packet)

Pat is forty four year old woman and grandmother. Her two daughter are drug addicted. Her daughters take the grandchildren the drug dealers farm. She is very worry and wants them to quit it.

In the parking lot, she was met by the police. The police wanted her to undercover in Gitfale Jimmy's gang. It is very dangerous, but Pat went Jimmy's house with police wire.

Her daughter introduce Jimmy to Pat. Pat said her friend wanted one pound of the drug for fifteen thousand dollars. Jimmy said he does not have it, because he does not have the special cemichal. So Pat asked the police for the cemichal and brought it to Jimmy.

In three weeks, Jimmy is trusted Pat. They went together to the lab. Jimmy was joking about the wire and Pat did not understand it was joking, so she was nervons. In the next day, Jimmy's gang was arrested by police, and they went to the jail. Pat's daughters are now free for the drugs.

Step 2 Questionnaire

Please do not open this questionnaire until you have completed Step 1.

Group _____

Code: _____

QUESTIONNAIRE

Personal Information

1. Age (circle one)

20-24 26-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-70

2. Sex M _____ F _____

3. First language(s) at birth and still used:_____

Other languages spoken or learned: _____

_____

4. Have you ever been a resident in a non-English speaking country? YES NO
If yes, please specify the countries and length of time:

_____

_____

Teaching Experience

1. How many years have you been teaching ESL? _____

2. How many of those years have you been teaching adults? _____

3. How many years have you taught adults in a multilingual classroom? _____

4. Have you taught adults in a monolingual classroom? YES NO
If yes, please specify how many years and the first language of the students:

_____

_____

5. Have you taught in a public school system in Canada? YES NO
If yes, please specify how many years and the grade(s):

_____

_____

6. Have you taught ESL courses where the specific focus was writing? YES NO
If yes, please specify the approximate number of courses you have taught:

_____

7. Have you taught ESL overseas? YES NO
If yes, please indicate the countries and length of time:

_____

_____

8. Most of your teaching experience is with (choose one):

_____ Low levels
_____ Intermediate levels
_____ Advanced levels
_____ All levels equally

9. Which levels do you prefer to teach? (choose one)

_____ Low levels
_____ Intermediate levels
_____ Advanced levels
_____ I don't have a preference

Experience at THIS INSTITUTION

Circle how many times you have taught these courses at ***THIS INSTITUTION*** in the past five years:

Intensive 1	0	1	2	3	4	5	6 or more
Intensive 2	0	1	2	3	4	5	6 or more
Intensive 3	0	1	2	3	4	5	6 or more
Intensive 4	0	1	2	3	4	5	6 or more
Intensive 5	0	1	2	3	4	5	6 or more
Intensive 6	0	1	2	3	4	5	6 or more
Intensive 7	0	1	2	3	4	5	6 or more
Intensive 8	0	1	2	3	4	5	6 or more
Writing 1	0	1	2	3	4	5	6 or more
Writing 2	0	1	2	3	4	5	6 or more
Writing 3	0	1	2	3	4	5	6 or more
Writing 4	0	1	2	3	4	5	6 or more
Standard Test Preparation	0	1	2	3	4	5	6 or more
Writing Test Preparation	0	1	2	3	4	5	6 or more
Interactive Grammar	0	1	2	3	4	5	6 or more
Study Skills	0	1	2	3	4	5	6 or more
Conversation 1	0	1	2	3	4	5	6 or more
Conversation 2	0	1	2	3	4	5	6 or more
Conversation 3	0	1	2	3	4	5	6 or more
Conversation 4	0	1	2	3	4	5	6 or more
Conversation 5	0	1	2	3	4	5	6 or more
Conversation 6	0	1	2	3	4	5	6 or more
Conversation 7	0	1	2	3	4	5	6 or more
Conversation 8	0	1	2	3	4	5	6 or more

Criteria for Evaluating Writing

Think about the level you just assigned to the writing samples. At that level, there would normally be some variation in the types of errors or aspects of writing that cause difficulties for different students. Some errors or problems might be considered hallmarks of a particular level and may be indicators for placement. Consequently, some types of problems or errors might be considered more "serious" than others at a level. In other words, at one level, basic word order errors may be considered much more "serious" than paragraph cohesion errors. At another, problems with fragment sentences could be expected to be behind the student and therefore constitute a serious error that the student should attend to.

The level you indicated for the writing samples _____.

Keeping in mind the level of the samples you have just evaluated, please rate the following aspects of writing in terms of how serious errors or difficulties with them are **for that level**.

A rating of 1 means you consider an error or problem with this aspect to be very serious at the level you have in mind and a rating of 6 is not serious.

	Very Serious					Not Serious
Articles	1	2	3	4	5	6
Capitalization	1	2	3	4	5	6
Fragment sentences	1	2	3	4	5	6
Handwriting	1	2	3	4	5	6
Organization	1	2	3	4	5	6
Paragraph cohesion	1	2	3	4	5	6
Plurals	1	2	3	4	5	6
Punctuation	1	2	3	4	5	6
Run-on sentences	1	2	3	4	5	6
Spelling of common words	1	2	3	4	5	6
Spelling of difficult words	1	2	3	4	5	6
Subject-verb agreement	1	2	3	4	5	6
Topic sentences	1	2	3	4	5	6
Using the wrong preposition	1	2	3	4	5	6
Using the wrong pronoun	1	2	3	4	5	6
Word order	1	2	3	4	5	6

APPENDIX D-11: Step 3 Instructions

Step 3

Instructions

Please answer the three questions below on the spaces provided on the back of each writing sample. Do not write your answers on this instruction sheet.

1. What do you think is the first language of this writer?

2. How certain are you that this is the first language of the writer?

3. What is it about this sample of writing that made you think that this is the writer's first language? (It is not necessary to do a full linguistic analysis of the sample: one or two features of the writing will do.)

Thank you very much for your help in the completion of this study. Your time and patience are greatly appreciated. Please don't discuss any part of this data collection session with anyone until all the data collection sessions are completed.

APPENDIX D-12: Blanks that Appeared on the Back of Each Sample

(for L1 Identification in Step 3)

1. _____

2.

_____ Very Certain
_____ Somewhat Certain
_____ Somewhat Uncertain
_____ Very Uncertain

3. _____

_____

_____

_____

Code: _____