Measuring rater consistency: An investigation into the effects of two testing instruments on raters' scores

Marcello Quintieri

A Thesis

in the

Department of Education

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts (Applied Linguistics) at
Concordia University
Montreal, Quebec, Canada

December 2005

Your file Votre référence
ISBN: 0-494-14170-0
Our file Notre référence
ISBN: 0-494-14170-0

ABSTRACT

Measuring rater consistency: An investigation into the effects of two testing instruments on raters' scores

Marcello Quintieri

Attaining rater consistency in evaluating compositions is both an important and elusive goal. The literature is filled with proposals and discussions on how to improve rater consistency. One approach that teachers at the institution where this study took place intuitively feel would promote consistency is the use of model compositions against which raters could measure the target compositions. This actual contribution to the literature, though, has never been tested.

This study investigates the effects of preselected model compositions and a multiple weighted trait scale on the consistency of raters' scores in a direct ESL writing assessment context. Twenty experienced ESL teachers at the institution where this study took place were divided into two equal groups. One group used model compositions and the weighted multiple trait scale to grade a set of sixteen authentic ESL exam papers. The second group graded the same set of compositions using the trait scale alone. Ratings were analyzed using traditional statistical techniques, such as a *t*-test analysis, as well as Minifac Facets software, a student version of the multi-faceted Rasch analysis program FACETS. Findings from the *t*-test showed more significant score variance among the teachers who used the trait scale alone to grade the compositions than those who used model compositions in addition to the trait scale. This difference between the two groups was supported in the Rasch analysis, which identified a greater spread in the over-all severity measures for the group of teachers who graded with the trait scale alone. In

addition, the Rasch analysis identified this same group of teachers as grading the compositions either most severely or least severely. These results suggest that the use of model compositions in addition to a trait scale improves interrater consistency.

To Mom and Dad

ACKNOWLEDGMENTS[1]

First, I would like to thank my supervisor Dr. Elizabeth Gatbonton for her vision and direction throughout all the stages of this thesis. I am also grateful to my readers Dr. Pavel Trofimovitch and Dr. Marlise Horst for their insightful comments and words of encouragement. All three of them have greatly contributed to what this thesis has come to be.

Second, I am indebted to the expertise of Dr. Vipahvee Vongpumivitch and Randall Halter and their helpful comments over several stages of this project.

Third, I would like to thank the teachers who helped me establish model compositions in the preliminary stage of the project: Scott Chlopan, Laura Cowan, Karla Holmes, Hagop Kassabian, Mike Miao and Nina Padden. I am also grateful to those teachers who participated in the actual study.

Fourth, I would like to thank Mike Padua and Dr. Everett V. Smith for their help with the Rasch analysis.

Last and certainly not least, I would like to name friends who supported me throughout this project: Dean, Hannah, Kerry and Sandra. Words cannot express my gratitude.

[1] Support for this thesis project came from funds allocated to Dr. Elizabeth Gatbonton from an FQRSC team grant, awarded, September 2005, to Norman Segalowitz, Principal Investigator, and collaborators, Michael Von Gruneau, Elizabeth Gatbonton, and Roberto Alameida.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1: Introduction

Theoretical work on the notion of communicative competence has influenced the way in which second language is evaluated. Within this communicative language testing framework, testing specialists and language teachers alike have applied various types of testing procedures to evaluate linguistic ability. For the most part, these testing procedures can be seen as possessing an "indirectness", or a characteristic that is "indirect", because they are designed to make inferences more about test takers' underlying competence than of their performance at the time of the test (Bachman, 1990; McNamara, 1996). Nonetheless, there are also language tests within this framework (i.e., that set out to evaluate a test taker's language competence) that have been referred to as "direct" tests because they engage the test takers to produce language that is more natural. Thus, communicative language testing can be viewed more as a continuum than as a strict dichotomy. At one extreme there are the more "indirect" tests, which involve assessment procedures that set out to measure test takers' language competence in less natural, more fixed-response language use contexts; that is, these are tests that utilize a more contrived measurement of language ability because they do not take account the realistic situations in which language is used. These more "indirect" tests are manifested in such procedures as multiple choice tests of grammar and usage, synonym-matching tests, etc. These tests can also be said to be mainly objective in nature because of the manner in which they are scored: the person scoring the test follows an established set of responses that do not require interpretation of language ability on the part of the person marking the test. For example, in multiple choice tests those marking follow an answer key. At the other end of the continuum are the more "direct" forms of testing, which involve the

inference of communicative competence in less contrived and more "actual" or "real-life" language performance situations. Examples of such tests are participating in an oral interview or writing an essay. These more "direct" tests can also be characterized as having test takers respond to more open-ended prompts, requiring potentially more creative and unpredictable language. While the purpose of these more "direct" tests is to elicit language that is more natural and meaningful than the language elicited in the more "indirect" tests, the judgements of test takers' communicative performance is more subjective, varied and less predictable. As a result, language testing specialists have set out to identify more reliable and valid measurement procedures to evaluate the less fixed-response, more open-ended nature of these so-called "direct" tests.

The reliability of ratings of learners' performance is an area of particular concern for testing specialists. Reliability, in a general sense, refers to the extent to which a test measures a candidate's performance in a relatively consistent and error-free manner. In a communicative language assessment context, Bachman (1990) discusses factors that are not related to the language ability of the test takers but still affect test scores. For example, in an ideal testing scenario where raters judge essays for a final exam, it would be assumed that upon following a set of criteria they would arrive at the same assessment of a test taker's performance. Any variation in their judgements would be construed as measurement error because the different scores interfere with the ability to make an accurate inference of the test taker's true language ability. Bachman refers to these factors as *potential sources of measurement error*, and he categorizes them into three mutually exclusive groups: 1. test method factors (e.g., raters, prompt type, etc.), 2. personal attributes (e.g., test-taker's cognitive style, knowledge of particular content,

etc.), and 3. random factors (e.g., fatigue, time of day, etc.). The objective in developing and administering "direct" tests, is to minimize these potential sources of error, or sources that cause score variance, in order to maximize reliability.

The first of Bachman's three categories, test method factors, constitutes the *how* of language testing, and the reliability of raters' judgements falls within this category. Striving to reduce these sources of error (such as rater variation) is particularly important because testing specialists have come to understand reliability as an element contributing to test validity, and that improving reliability satisfies a necessary condition for validity (Bachman, 1990). While some potential sources of error, such as rater fatigue, may be unpredictable, other sources, for example, the severity of raters' judgements, may be systematic—that is, some raters may be consistently, and therefore predictably, more severe in their ratings than other raters. Both these random and predictable factors that cause variation in raters' judgements can occur over time (i.e., intra-rater consistency) and across raters (i.e., inter-rater consistency). Moss (1994) considers rater consistency as the greatest problem to overcome in the field of second language assessment because it is concerned with reconciling inherent subjectivity and the need for objective precision.

Inherent in "direct" second language testing, then, is the element of rater subjectivity—that is, raters may interpret assessment criteria differently and, for example, be particularly harsh in their judgements towards grammatical errors, while others may be fairly lenient. This will lead to score variance and obscure inferences of test takers' true language ability. It is important to minimize raters' inconsistencies by seeking ways to bring objectivity to their judgements. The purpose of this study, then, is to investigate the effects of a rating procedure that may enhance rater consistency in this way.

The thesis project to be reported here focuses on rater consistency in evaluating written material. In particular, it examines the issue concerned with ensuring that raters evaluating the same set of compositions arrive at more or less similar judgements about these materials. Two ways of ensuring greater rater consistency are investigated in order to find out whether and how they contribute to the attainment of this goal.

Rationale

Rater consistency in first language (L1) and second language (L2) large-scale direct writing assessment studies has been investigated from several perspectives and research on this issue can be categorized as follows: 1. studies considering factors that influence rater variance, 2. studies considering ways to control rater variance, and 3. studies viewing rater consistency either as rater reliability or as rater severity. Studies that have investigated factors that influence rater variance have considered elements of writing attended to by raters, prompt-type, duration of time, and raters' backgrounds. Studies concerned with the second perspective either have investigated ways to bring raters into agreement on an established set of standards or have investigated the usefulness of the manner in which these standards have been objectified. Moreover, research that has investigated ways to control rater variance has interpreted rater consistency either as rater reliability or as rater severity. For instance, studies implementing traditional statistical procedures (e.g., compute coefficient alpha to determine inter-rater reliability) have been concerned with analyzing rater agreement from a more general perspective—one that assumes measurement error as being random. This is a necessary assumption for statistical procedures used to measure potential

sources of error within the classical 'true' score measurement model.[1] Other studies have adopted a perspective that can be viewed as more specific because the statistical tools applied identify measurement error that is systematic. The ability to identify such systematic error is important because it allows for a greater understanding of factors affecting rater variance, thereby leading to an improvement in rater consistency. What follows is an overview, albeit not an exhaustive one, of research from the three perspectives mentioned above.

Studies Investigating Factors that Influence Rater Variance

An early paper that raised the concern about uniform rater scoring was Diederich, French and Carlton (1961) who found considerable variation in the judgements of 53 raters using a 9-point scale to grade 300 writing test samples produced by native speakers of English. The purpose of the study was to reveal differences of opinion among readers of student writing samples for college admission. The study reported .31 as the mean correlation between raters. In addition to raters agreeing with each other only three times out of ten, a minimum of seven different grades were assigned to 94% of the papers and no paper received less than five different grades. The study also categorized over 11,000 rater comments and identified five basic features of writing attended to by raters, with *IDEAS* emerging as the most important feature. The other categories were *FORM* (i.e., organization), *FLAVOUR* (i.e., style), *MECHANICS* (i.e., errors in grammar and punctuation), and *WORDING* (i.e., choice and arrangement of words).

Subsequent studies investigated factors influencing score variance in direct writing assessment from various perspectives. Huot (1990) cites numerous studies that

investigated the extent to which elements of writing, such as grammar and content, influence raters' perceptions. While not unanimous, the majority of these studies indicate that content and organization were more influential than mechanics and sentence structure when raters evaluated writing quality in L1 contexts. Surprisingly, relatively few, if any, L2 direct writing assessment studies have investigated causes of rater variance from this point of view.

Research in L2 writing assessment, however, has also reported that prompt type may be a source of rater variance. Weigle (1998), for example, found a significant difference in the manner in which two essays based on two different prompt types were evaluated by a group of inexperienced, untrained raters. One prompt required a response to a graph, while the second required a response to information presented in a table. Weigle (2002) suggested that the rater variance may be related to the tendency of one prompt to elicit several rhetorical patterns unlike the latter, which elicited traditional five-paragraph essays. While further investigation would shed more light on the extent to and reasons for which prompt type influences raters' perceptions, research in this field nonetheless indicates that raters' judgements can vary across task types.

Furthermore, rater variance has been shown to occur over time. In an L1 context, McQueen and Congdon (1997), for example, investigated the scores for 8,285 elementary students' writing performances over a 7-day rating period. Results of the study showed that in spite of monitoring, significant differences in rater severity persisted for each day as well as in the 7 days combined. In addition, these changes in severity could produce differences of more than half a score point on a six-point holistic scale. In addition, raters tended to grade the same papers more severely on the second occasion. McQueen and

Congdon's findings confirm concerns raised by other similar studies, such as Lumley and McNamara (1995), which found significant changes in rater severity for three sets of rating over a 20-month period for a test of spoken English.

With respect to the influence of raters' different backgrounds on score variance, Weigle (1998) also reported significant rater variance between inexperienced (those who were not familiar with the rating scale) and experienced (those who were familiar with the rating scale) raters in the pre-norming session—that is, before training. The study also found that inexperienced raters tended to give more extreme scores and applied the rating scale more severely. In the post-norming session, however, no clear distinctions between inexperienced and experienced raters were found. Shohamy, Gordon and Kraemer (1992), also investigating the effects of training, similarly reported that background (i.e., experienced v.s. non-experienced raters) did not correlate with raters' scores. Thus, the research on this issue appears to indicate that following training rater background differences do not contribute to score variance.

Studies Concerned with Ways to Control Rater Variance

Weigle (1998) and Shohamy et al. (1992) are also important because both represent the few studies that have investigated the effects of training on rater consistency in ESL writing contexts. The training of raters is important because it is a way to bring raters into agreement on a set of criteria, thereby bringing more objectivity to raters' judgements. Some studies, such as Lumley and McNamara (1995) and McQueen and Congdon (1997), question the practice of certifying raters on the basis of a once-only

calibration, and go so far as to advocate the constant monitoring of raters to keep score variance in check.

Two concerns, nonetheless, arise when considering research on rater training in a direct writing assessment context. First, while there is an attempt to bring raters into agreement, there is no established protocol that has been empirically proven to be more effective than any other. Therefore, training procedures across administrations can vary. Bachman & Palmer (1996), for example, outline a six-step protocol for training raters. Important components of this general procedure involve the review, application and discussion of: (1) the rating scale, and (2) the language samples, or model compositions, graded by expert raters. These two components found in rater training procedures are ways to establish objective standards. The second concern when considering research on rater training is that very little, if anything, is known about the utility of model compositions—only the usefulness of rating scales on rater variance has been empirically scrutinized.

For example, holistic scales, where raters give a single score to a writing sample, have been seen as more subjective than analytic scales, where raters calculate the sum of separate scores attributed to the different elements of writing, (e.g., content, organization, etc.). Studies comparing score variance using both scales have found that analytic scales produce higher levels of rater consistency (Weir, 1990; Bacha, 2001). This may be the case because analytic scales further objectify the assessment criteria by presenting components of written language as five main parts. In addition, their use causes raters to focus their judgements on these language components.

Still, in spite of the use of an analytic scale rater variance is not eliminated. One reason for this is given by Vaughan (1987; cited in Huot, 1990) who noted that raters tend to rely on their "individual judgements" when papers do not conform to specific criteria on the rating scale. Lumley (2002), focusing on assessing L2 compositions, similarly suggests that raters use their "complex intuitive impressions" when such situations arise. Therefore, when elements of grammar or organization, for example, found in the writing sample are not found in the rating scale, there is more room for interpretation, which may lead to greater variation in rater scoring. Turner and Upshur's (2002) study suggests that rater consistency can be enhanced when raters work together to develop rating scales that are based on the characteristic features identified in the written test samples. Moreover, the objectified standards described on the scale are interpreted more similarly because raters have worked together to develop the rating scale.

The second component found in rater training procedures—that is, language samples, or model compositions—are used to help raters apply the criteria on the rating scale when evaluating the test samples. As they have been graded by expert raters, they represent examples of, for instance, an A paper, B paper, etc., to the rater trainees. Model compositions, therefore, can be seen as an attempt to objectify standards because they include those same characteristic features found in the test samples and described in the rating scales. This, thereby, helps the raters interpret the criteria found on the rating scale more similarly. Research on the utility of model compositions in direct ESL writing assessment contexts, however, has been neglected. Their value, nonetheless, has been investigated in other contexts. Dandenault (1997, unpublished M.A. Thesis), for example, looked at the usefulness of voice samples to help place adult learners in the

English as a Second Language (ESL) classroom. Her study adapted a self-assessment tool used in Segalowitz (1976), whose study indicates that the creation of a standard with which learners could compare themselves allowed for better prediction than a self-assessment procedure using a 7-point Likert-type scale alone. In Dandenault's study, a group of French speaking ESL students rated themselves against selected voice samples representing approximately the 2, 4 and 6 categories of a 7-point scale. Her study's results suggest that the use of these models brings about satisfactory levels of rater consistency. As Dandenault's study has investigated the effectiveness of models as a self-assessment tool to evaluate speaking ability, it would be worthwhile investigating their value on rater behaviour in a direct ESL writing assessment context. The purpose of this study, then, is to investigate the usefulness of grading written texts with model compositions, in addition to an analytic rating scale, towards improving rater consistency.

Understanding Rater Consistency

In addition to investigating the effects of rater training, Shohamy et al. (1992) and Weigle (1998) are important because they represent a third perspective in the literature by offering two ways to understanding rater consistency—that is, as rater reliability and as rater severity. The reason for this lies in the way in which each study applies statistical procedures to analyze data. Shohamy et al. (1992), using an intraclass correlation analysis, represents those studies that investigate "rater reliability"—that is, those that apply statistical procedures that make the assumption that potential sources of error (e.g., rater variance) are random and unsystematic. On the other hand, the study by Weigle (1998), which investigated "rater severity", represents studies that use statistical

procedures that identify systematic factors contributing to rater variance (e.g., the severity of raters' judgements). While both studies approach rater consistency from different perspectives, it would be worthwhile determining whether the two are at all complementary—that is, there will be an attempt to show that one statistical procedure can be used to support the findings of the other. Therefore, in addition to investigating the usefulness of model compositions on rater variation, this study will analyse the data using two different statistical tools: one that is used to determine consistency with respect to the reliability of raters' scores, and the other that is used to measure the severity of raters' judgments.

Purpose of the Study

To summarize, the first goal of this study is to investigate the extent to which two testing instruments (i.e., models and rating scale) affect rater reliability in a direct ESL writing assessment context. Models refer to authentic compositions that serve as benchmarks against which raters grade the target compositions. Rating scales represent the analytic grids that raters also use to grade these compositions. A second purpose to this study is to determine how a combination of models and the rating scale influence rater severity. Since no study to date has considered the effect of model compositions with respect to rater reliability and rater severity in a direct ESL writing assessment context, studies that have investigated the effects of rating scales and rater training on score variance will be reviewed. What follows in the next section, then, is the context for such an investigation.

Chapter Two: Review of the Literature

This chapter will review studies that have investigated rater consistency. First, studies that have viewed rater consistency as rater reliability will be considered. From this perspective, research on the usefulness of rating scales and the effects of rater training will be noted. Second, discussion will turn to research that has interpreted rater consistency as rater severity. The contribution this perspective makes to the literature as well as the effects of rater training on rater severity will be discussed. Finally, the importance of investigating the usefulness of model compositions on rater consistency will be made.

Rater Reliability

Studies that are concerned with the reliability of ratings view rater consistency from a more general perspective, interpreting measurement error as random and homogenous. In addition, the statistical analyses of these studies are concerned with the relative degree of rater agreement.

Gamaroff's (2000) study is relevant because of the conclusions it makes. The study investigated the variation of scores given by six groups of four raters assessing 50 high school entrance exam compositions written by two groups of ESL learners. The raters were experienced teachers at various universities, technikons (i.e., technical institutions) and colleges in South Africa, and the rating session took place at a language assessment workshop conducted at a National Association of Educators of Teachers of English (NAETE) conference in South Africa. The raters used a 9-point holistic scale to evaluate two types of essays. They were also requested to work individually, spend 1 ½

minutes on each composition, give an impressionistic score based on content, grammatical accuracy, and any other writing feature they wished to consider, and provide reasons for their judgments. Data were analyzed in three ways: 1) by comparing the scores of each rater, 2) by comparing the average scores of each group of raters, and 3) by comparing individual rater's judgments with their scores. Findings showed that in spite of rater experience, there was substantial variation in the elements of writing attended to by the raters. For example, rater judgments on one of the prompts were 54% negative towards content and 42% negative towards grammar. Furthermore, there was disagreement among some raters as to how to categorize certain errors. For example, what some raters interpreted as a grammatical problem, others saw as a lexical problem. Gamaroff concluded that objective standards and agreement on those standards are necessary if direct writing assessment is ever going to produce high levels of rater consistency.

One way to obtain objective standards is to use rating scales. Weir (1990) reported differences in holistic versus analytic scoring, where the former involves giving a single score to a writing sample, and the latter involves calculating the sum of separate scores given to the different performance dimensions (e.g., content, organization, etc.). Weir's study reported that holistic scoring provided less reliable results. Similarly, Vaughan (1991) indicated that holistic grading can be highly subjective, and that scores within a rater and between raters can vary significantly. The author concluded that analytic scoring would provide a more objective form of direct assessment. Hamp-Lyons (1991), similarly arguing for a more objective assessment procedure, suggests that holistic assessment may only be appropriate for specific contexts and that, more often

than not, this assessment procedure neglects the internal complexity of student writing samples.

At least one study directly compared holistic and analytic scales in terms of resulting rater consistency. In her study, for example, Bacha (2001) asked experienced raters to judge compositions using either holistic or analytic evaluation criteria. The study used a stratified random sample of 30 essays from a corpus of 150 final exams written by L1 Arabic students. The exams followed a four-month ESL reading and writing course, which was the first of four levels given at the American University of Beirut. The exam had students write an essay in response to one of two prompts, where each prompt required a different rhetorical pattern studied during the term. Two experienced raters, the class teacher and a teacher of the same course but of a different section, graded the compositions first, using a holistic scale and then, using a multiple-weighted analytic grid with five components: content (30%), organization (20%), vocabulary (20%), language (25%), and mechanics (5%). The two raters based their judgements on the same final course objectives profile when applying the two grading systems. The study reported slightly higher correlation coefficients for the analytic scale (.80+); nonetheless, the holistic scale did produce correlation coefficients of .80. That the two raters were experienced and very familiar with the course objectives may have contributed to the high levels of rater consistency in both assessment procedures. Still, the research indicates the superiority of analytic rating.

The concern for bringing raters into agreement on established objective standards can be observed in studies investigating rater training. Shohamy et al. (1992) investigated the effects of training on two groups of raters: 1) those with experience in

teaching ESL writing, and 2) those with no teaching experience but identified as native speakers of English. Four groups of five raters graded 50 test samples using three versions of a holistic grid. The first grid focused on general writing quality, the second assessed content, and the third focused on vocabulary and grammatical structures. Two groups, one consisting of experienced raters and the other inexperienced raters, received training before the grading of the compositions. The two other groups, also divided into experienced and inexperienced raters, received no training. Findings from the analyses using Ebel intraclass correlation, a formula that computes a coefficient alpha from the sums of different raters' ratings, showed high correlation coefficients for all groups, ranging from .80 to .93. However, trained raters regardless of background and the version of holistic scale used had interrater reliability scores ranging from .91 to .93. Finally, the Spearman-Brown correlation formula, which was used to evaluate the internal consistency of the raters' judgements, yielded intra-rater reliability scores of trained raters following an interval of three weeks ranging from .76 to .96 for all three scales. These findings show that while background does not influence the reliability of rater scores, training does have a significant effect on rater reliability. The authors also note that the empirically based rating scales—that is, rating scales that were based on the characteristic features identified in the written samples—may have contributed to the high levels of rater reliability.

The effects of training on trained and untrained raters was also investigtated by Weigle (1994). Unlike Shohamy et al., this study used a pre/post design and think-aloud protocols, which involved recording the verbalized thoughts of raters while they scored the ESL compositions. Raters also graded compositions with an analytic scale. The

training process lasted two hours and involved comparing, discussing and understanding the rationale behind official scores for compositions that spanned the different bands of the rating scale. Weigle found that marked differences (i.e., raters whose scores between the pre and post ratings differed by three points or more) in the pre/post training scores occurred in 50% of the new raters, but that there was no distinct pre/post training-score difference for the other half of the new raters nor for all of the old raters. In addition, the post scores of the new raters were closer to the group mean than were their pre-scores. From these findings, Weigle concluded that training very likely had a positive effect on the variability of raters' scores because it helped them understand the criteria of the grid better. Weigle arrived at this conclusion by making a simple comparison of the score differences between the pre and post ratings for each rater and by supporting these findings with the raters' verbalized thoughts obtained from the think-aloud protocols. Weigle (1994), unlike Shohamy et al. (1992), did not go so far as to calculate correlation coefficients for the separate groups of raters to determine any effects on rater reliability. Still, both studies indicate that rater training has a positive effect on score variance.

Rater Severity

Studies that investigate the severity of ratings measure the degree to which raters' scores are not equal. In addition, these studies view rater variance as resulting from sources of measurement error that are in part systematic.

So far, the studies discussed in the literature represent a common view regarding the issue of rater variation. But before moving on to a discussion of the perspective shared by these studies, it is necessary to return to the issue of measurement error

associated with language testing. As Bachman (1990) pointed out, sources of error that are not associated with the abilities of the test taker affect the reliability of the test. One factor that contributes to test reliability is referred to as test method facets, and within this category is rater behaviour. In other words, the judgement of raters constitute a potential source of measurement error (i.e., rater variance) that has an effect on language test scores. For example, in writing assessment, raters may focus on different components of writing such as content, organization, etc. They may also interpret the rating scale differently, so what is a "B" for one rater is a "C" for another. However, the studies that have been discussed up to now are not able to distinguish these different sources of rater error that may be systematic and interact with each other. The reason for this lies in understanding classical true score (CST) measurement theory and the models associated with this theory. CST essentially assumes that test takers' observed scores are the result of a true score (or their true ability) and an error score (or factors that affect their true score but are unrelated to their language ability—e.g., rater severity). The measurement models used to analyze the data—that is, the computation and interpretation derived from CST models, such as the coefficient alpha—can only view error score as random and homogenous (Bachman, 1990). In other words, CST models, while still useful, have limitations because they fail to break down and distinguish sources of error that may be systematic. There is no way of knowing, for example, how a particular rater judged a particular question or interpreted a particular step in the scale. Understanding these and other types of interactions provide a clearer understanding of the variance associated with raters' judgements.

Developments in measurement, such as Item Response Theory (IRT), have provided solutions to these limitations. While there exist many IRT models, a common feature is that they all attempt to draw conclusions about the underlying ability of a candidate and difficulty of a test item (Bachman, 1990; McNamara, 1996). One IRT model, which has been used in studies concerned with the variability of raters' judgements, is the Multi-Facet Rasch Measurement (MFRM) model developed by Linacre (1989). This model falls within the family of Rasch measurement models which calculate (*calibrate*) estimates (*measures*) that take into account characteristics of the test situation that can influence test scores. Early versions—that is, versions that predate the MFRM model—calibrate measures by taking into account two *facets*, or aspects of the test situation, that can influence test scores. Some of these earlier versions are the Rating Scale Model (Andrich, 1978), which has been used for dichotomously scored tests (e.g., true/false tests) and the Partial Credit Model (Wright & Masters, 1982), which has been used to handle polytomous data—that is, data obtained from tests requiring responses on Likert-type scales. These earlier models calibrate dichotomous or polytomous data by relying on a mathematical relationship between the ability of the test takers (*candidate ability*) and the difficulty of the questions (*item difficulty*). This mathematical relationship is expressed as the probability of a particular response.

Calculating probabilities about candidate ability and item difficulty is characteristic of all Rasch models.[2] The MFRM model contributes to the family of Rasch measurement tools by its ability to calibrate multiple facets (i.e., more than two) simultaneously. Rater behaviour, for example, is a facet, in addition to candidate ability and item difficulty, that influences test scores. The MFRM model, then, allows for the

estimation of the ability of a candidate on a given item by a given rater by taking into account the interaction of all candidates, items and raters in the data set simultaneously (Linacre, 1989; McNamara & Adams 1991/1994). Again, this is achieved by using a mathematical equation outlined in the Results chapter below. An objective of this study is therefore to detect and measure rater behaviour as rater severity based on the relative ability of the candidates and difficulty of the items. In addition to calibrating measures by taking into account multiple facets simultaneously, the MFRM model, as other Rasch models, expresses these measures individually and plots them onto a true interval scale, where the distances between intervals are equal. This type of scale is known as a *logit* scale because the probabilities are expressed as a logarithm. The advantage of expressing these measures as logits is that in addition to telling us, for example, that one rater is more severe than another rater, it can indicate by *how much* this particular rater is more severe. The measures on this scale can then be compared to other facets influencing test scores, such as candidate ability and item difficulty, as well as the steps on the rating scale used to grade a test taker's performance. These aspects can be easily compared as they are all measured against the logit scale. From this information, one can determine any problematic effects resulting from, for instance, the severity of raters' judgements. For example, Engelhard (1992) identified a significant spread (that is, a significant variability) in rater severity for the writing portion of a high stakes high school entrance exam. Findings of this study showed that where some candidates failed, others of equal ability passed, and this was true in spite of an interrater reliability coefficient of .82.

In addition to pinpointing which raters are more/less severe in their ratings and indicating by how much, the MFRM model can provide information on peculiar patterns

as well as systematic subpatterns of rater behaviour. This is possible through an analysis of the *fit statistics*, which is a set of statistics the Rasch model produces, in addition to the logit measures. The Rasch model arrives at these statistics by comparing the probabilistic (*expected*) scores to the raw (*observed*) values. Essentially, then, the Rasch model compares two sets of data—one based on the observed scores and the other based on the calibrated measures obtained using the logarithmic equation—to produce these fit statistics. From this information, the model can indicate, for example, that a rater is unexpectedly too severe on a particular item and that this may be the result of a certain interaction pattern (e.g., rater-step interaction, rater-ratee interaction, etc.). Thus, the Rasch model provides an explanation as to *why* raters are more severe in their rating for a given case. This precise and meaningful diagnostic information provided by the MFRM model has led to its increased use in studies investigating rater consistency and, in particular, rater training.

One recent example of the use of the MFRM model is found in Weigle's (1998) study of rater severity. More specifically, Weigle used the data from her 1994 study and the MFRM model to investigate rater severity. In the pre-norming session, data showed that both the new and old raters differed with respect to rater severity. More specifically, the new raters were more severe and gave more extreme scores. Furthermore, the think-aloud protocols of the new raters revealed that they applied the scoring rubric more rigidly than the old raters. In the post-norming session, data revealed a reduction in the spread of severity—that is, the scores were not as extreme as in the pre-norming scores. However, differences in rater severity were still found to be significant. In addition, intra-rater consistency improved for most but not all raters, and it improved considerably

for three of the new raters. Overall, however, no clear distinctions with respect to severity or consistency could be found between new and old raters. By applying the Rasch model, Weigle showed that training cannot eliminate interrater inconsistencies entirely, but it can bring about satisfactory levels of intra-rater consistency.

The findings of Weigle's study confirm those from previous and subsequent studies investigating the effects of training on rater severity. Studies such as Wigglesworth (1993), Lumley and McNamara (1995), McQueen and Congdon (1997) and Kondo-Brown (2002) all report that while training has the effect of making raters more self-consistent, it does not necessarily lead to raters evaluating performance equally severely. Nonetheless, where significant differences in rater severity persist, satisfactory levels of rater reliability are still possible. Furthermore, as has been shown in Engelhard (1992), Rasch analysis can reveal the existence of problematic rater effects even when rater reliability is high. Similarly, Kondo-Brown (2002), which found that raters produced highly correlated scores in spite of significant variation in rater severity, also reported the existence of rater-candidate interaction among those candidates whose ability was extremely high or low, suggesting that raters were biased towards these types of test takers.

Perhaps the best final word on training comes from Weigle (2002), who notes that studies have yet to show that raters will ever completely agree on writing scores; however, studies have shown that training can help bring raters to some agreement:

> Raters bring their own backgrounds, experiences, and values to the assessment of writing, and while training can help bring raters to a temporary agreement on a set of standards research has consistently shown that raters will never be in complete agreement on writing scores (p. 72).

To summarize studies investigating rater reliability, findings show that training can bring raters into relative agreement on established objective standards. In studies investigating rater severity, findings indicate that exact agreement on established standards cannot be attained though acceptable measures of internal consistency is possible.

Finally, this review of the literature shows that studies have considered two general ways to bring more consistency to direct writing assessment. Some studies have considered ways to bring raters into closer agreement on established standards. Other studies have investigated the effectiveness of certain testing instruments (e.g., holistic and analytic rating scales) that objectify standards as a way of improving rater consistency. Rater training can be viewed as a common procedure for the former. Typically, included in rater training procedures is a review, application and discussion of the rating scale and compositions pre-graded by expert raters—that is, raters are trained to use two testing instruments that represent objectified standards. While studies have investigated the effects of holistic and analytic rating scales on rater consistency, there is one other means of bringing greater rater consistency that has not to date been examined—this is the use of pre-selected model compositions.

Pre-selected model compositions are, in effect, compositions that are graded by expert raters and are used in rater training procedures. These models contain elements that exemplify what are expected for each letter band, or step, found on the rating scale. In other words, model compositions are used as examples of an A paper, B paper, etc., and are intended for the purpose of helping raters consistently interpret established objective standards that are described on the rating scale. Research has indicated that analytic scales—that is, rating scales that show elements of writing which are broken

down into its main components (e.g., content, organization, etc.)—objectifies standards and helps raters grade more consistently. As there is no research in direct writing assessment that indicates the usefulness of model compositions, it is only assumed that they help raters consistently discriminate the different grades that can be attributed to the test takers.

Research on model compositions is, therefore, important because it will bring to light the utility of this testing instrument in rater training procedures. Moreover, understanding to what effect models contribute to rater consistency may lead to better procedures—that is, procedures that are more useful, less time consuming and more cost effective. Dandenault's (1997, in her unpublished M.A. Thesis) study reported that models can be a useful self-assessment tool to evaluate speaking ability. In direct writing assessment, studies suggest that rater consistency can be enhanced through the use of empirically-based rating scales—that is, rating scales that contain features of writing that are characteristic of those found in the evaluated samples (Shohamy et al., 1992; Turner & Upshur, 2002). It would be worthwhile, then, to investigate the utility of authentic exam papers selected as benchmarks against which raters evaluated target compositions.

The purpose of this study is to compare how two composition evaluation instruments affect raters' scores. More specifically, the study investigates the extent to which the use of pre-selected model compositions (henceforth, Models), in this case compositions representing various levels of academic English writing ability, and/or the use of a weighted multiple-trait scale (henceforth, Trait Scale), has an influence on score variance.

The research questions are the following:

1. Does grading student compositions with a Trait Scale + Models bring about the same overall level of rater reliability as grading with a Trait Scale alone?

2. Does grading student compositions with a Trait Scale + Models bring about the same spread of rater severity as grading with a Trait Scale alone?

These questions were investigated by comparing two groups of ESL teachers who were asked to rate the compositions of an authentic group of ESL students, using either the Trait Scale alone or the Trait Scale plus Models.

Chapter Three: Methodology

This chapter presents the various elements and phases involved in designing the present study and collecting the data. These are presented in the following order: Participants, Materials and Data Gathering Procedure.

Participants

Twenty ESL teachers recruited from a local university served as the participants (henceforth, raters) in this study. All raters had a minimum of three years experience teaching ESL with at least one year experience teaching ESL writing. All had had previous exposure to one of the rating instruments used in the present study (the Trait Scale) but no exposure at all to the other rating instrument (the Models). These teachers were divided into two comparable groups after they had completed a biographical data questionnaire (Background Questionnaire—see below) and after they had evaluated two practice compositions each (henceforth, Practice Compositions—see below). Matching was in terms of the extent to which the raters had taught ESL courses being offered at the university as well as the experience they had teaching and evaluating compositions. They were also matched in terms of the scores they gave to the practice compositions such that raters who gave a similar range of scores were represented in each group. For example, if four raters gave the two Practice Compositions the letter grade "B", two raters were assigned to one group and the two others were assigned to the other group.

In terms of language, thirteen raters specified English as their first language, six raters specified another language other than English as their first language and one teacher indicated two languages (French and English) as his L1. All had nativelike or

near nativelike facility with English. All raters indicated an ability to speak and/or write at least one other language. In terms of educational background, all the raters have an M.A. in Applied Linguistics or a related field, and three hold a Ph.D. in literature or education. The average age of the raters at the time of the study was 49 years. Thirteen of the twenty raters were female.

Materials

The materials used in this study included three questionnaires (a biographical data questionnaire, a debriefing questionnaire and a questionnaire to find out if the raters read the Models). The materials also included two rating instruments (Trait Scale and Models), two sets of written instructions on the use of these instruments and two types of compositions (Practice Compositions and Target Compositions) evaluated by the raters.

1. Questionnaires

The first questionnaire (the Background Questionnaire) was designed to seek information about the participants' age, gender, educational background, teaching experience, and experience rating compositions. One purpose of the Background Questionnaire (as mentioned earlier) was to help the researcher divide the participants into two equal groups according to characteristics given above. A copy of this questionnaire is presented in Appendix A. The purpose of the debriefing questionnaire was to determine whether the raters used their testing instruments (i.e., Trait Scale and Models) when evaluating the compositions and to what extent they believed these instruments were helpful. Since one group used the Models in addition to the Trait Scale,

their questionnaire contained additional questions focusing on the use of the Models. Copies of these questionnaires are presented in Appendix B (for the Trait Scale group) and Appendix C (for the Trait Scale + Models group). The questionnaire used to determine whether the raters read the Models consisted of two pages of questions that elicited general comprehension of the contents of each Model. A copy of this questionnaire entitled, *Pre-grading Questionnaire*, is found in Appendix D.

2. Rating instruments

Trait Scale: Participants were asked to use a weighted multiple-trait scale (Trait Scale), also known as an analytic grid, with accompanying band descriptors in rating the compositions. The particular Trait Scale offered to the teachers was one that is in current use for evaluating students in the advanced course (ESL 209) in the department where the teachers were recruited. The Trait Scale was designed to have five performance dimensions, with each dimension weighted differently: content (15%), organization (15%), grammar & language use (50%), vocabulary (15%), and mechanics (5%). The different weighting was based on the relative importance of each performance dimension in the stated course work objectives. Student performance on each dimension was graded on a 13-point scale using letter grades ranging from A+ to D- or F (no E marks were assigned). Each letter grade has a corresponding numerical value, which varied according to the dimension being evaluated. Thus, if content was evaluated with an A+, the corresponding numerical value was 15. The same letter grade in grammar & language use had the numerical value of 50. Each composition received a score for each

performance dimension as well as a total score that was the sum of the five dimensions. A perfect total score was 100. The Trait Scale is presented in Appendix E.

The Trait Scale also included a set of descriptors indicating discrete points on which the raters could base their scoring judgements (e.g., content included *Thesis statement*, *Introduction*, *Topic development*, etc.) A strong *Thesis statement*, for example, was one that was judged to be very clear and appropriate, while a weak *Thesis statement* was one that was unclear or not apparent. Raters were asked to check *strong* points and circled *weak* ones. Accompanying the Trait Scale was a two-page band description outlining the varying degrees of strengths and weaknesses for the discrete points. Essentially, then, the two-page band descriptors describe what the thirteen letter grades (i.e., steps in the thirteen-point scale) mean in each of the five performance dimensions. The band descriptors are presented in Appendix F.

Models: The Models were five authentic final exam papers selected to serve as benchmarks, or models, against which the Target Compositions could be marked. Each Model represented one of the following: 1) an excellent paper (deserving an A grade), 2) a very good paper (B), 3) an average paper (C), 4) a below average paper (D), and 5) a failing paper (F). The selection of these Models is described in the Data Gathering Procedures (see below). These five Models were argumentation essays that were written in response to the same prompt and were of the same length as the Practice and Target Compositions. None of the Models were used as Practice or Target Compositions. An example of a Model is presented in Appendix G.

3. Sets of written instructions on how to use the Trait Scale and Models

To ensure that the raters used the Trait Scale and the Models properly, sets of written instructions were prepared for them. Instructions in using the Trait Scale included a directive urging the raters to take note of the letter grades and corresponding numerical values given for each performance dimension of the Trait Scale. A directive also instructed participants to read the band descriptors for each performance dimension. In addition, instructions encouraged raters to indicate the letter grade and corresponding numerical value for each performance dimension (e.g., A+ and 15 for content) for each composition they graded. Raters were also asked to identify the Trait Scale descriptors that featured in their evaluation of each dimension by checking the strong items and circling the weak ones. Raters were also encouraged to add any comments in the space provided to the right of these descriptors. Finally, instructions directed raters to calculate the overall score by summing up the numerical values for each performance dimension. The set of written instructions for the use of the Trait Scale is presented in Appendix H.

Instructions for the use of the Models included a sentence directing the raters to read each Model carefully in preparation for answering a few questions about its contents. Since the raters were very familiar with the Trait Scale, it was important that the Trait Scale + Models group were given additional questions to ensure that they read, understood and used the Models when grading the Target Compositions. Thus, before the raters began grading the Target Compositions, they were given the *Pre-grading Questionnaire* discussed above. Instructions for the use of the Models also asked the raters to choose, for each Target Composition evaluated, a Model (e.g., Model B) against which the paper could be compared. This was to be done not just for the overall grade

but also for the grade for each performance dimension. In other words, raters had to indicate whether on each dimension (e.g., content), the Target Composition was *better than, as good as,* or *worse than* the selected Model using a 7-point Likert-type scale and then to assign a corresponding grade to this dimension, using a letter grade, for example: (B), its minus (B-) and plus version (B+), as needed. The set of written instructions for the use of the Models is presented in Appendix I.

4. Compositions for Evaluation

Two sets of compositions were given to the raters to evaluate. The first set consisted of two Practice Compositions, which were given to all the participants at the beginning of the rating session. The second set consisted of sixteen Target Compositions that all the participants graded in the main part of the study.

The Practice Compositions were two argumentation essays selected from the archives of final exams kept by the Credit ESL program at the participating university. For the final exams, the most advanced group of students in the program (ESL 209 students) were required to write an essay of approximately 500 words in length in response to one of a set of argumentation prompts. These exams had been rated by one instructor and verified by a second instructor. The latter was the course instructor for the student whose composition was being graded. The former was an instructor who taught a different section of the same course during the same session. The instructors graded these essays using the same Trait Scale and band descriptors described above. The Practice Compositions for this study were two ESL 209 final exam essays written in response to the prompt: *Children should be disciplined with physical punishment. Agree*

or disagree. Only essays consistently marked by the ESL teachers to have a total score corresponding to a C grade were used because greater variation in rater judgements can be identified with a middle range C paper than a graded paper that is closer to top or bottom grades (A or F, respectively).

The purpose of the Practice Compositions was twofold. First, the participants used them to practise their use of the Trait Scale on and its accompanying band descriptors in evaluating the compositions. Second, the participants' evaluations of these compositions were compared to one another to determine differences in the scores. As mentioned, the Practice Composition scores were used to divide the participants in to two matched groups so that they were comparable to each other in terms of the spread of their score differences. The names of the students who wrote the Practice Compositions (or any of the compositions evaluated in this study) were not revealed to the participants. An example of a Practice Composition is presented is Appendix J.

The Target Compositions were 16 argumentative essays selected from the same bank of final exams as the Practice Compositions. These essays were, thus, of similar length and were written in response to the same prompt as the Practice Compositions. None of the Practice Compositions were used as the Target Compositions. The main criteria for selection of the Target Compositions was their having been assigned the same mark by two experienced ESL teachers—that is, each composition had been graded by one teacher and confirmed by a second teacher—and assigned the grade of A, B, C, D, or F following the system described above (see Rating Instruments). An example of a Target Composition is presented in Appendix K.

Data Gathering Procedure

The data gathering procedure for the present study involved three stages: the selection of the Models, the distribution of Practice Compositions, and the distribution of Target Compositions.

1. Selection of the Models

The selection of the Models involved fifteen teachers with a minimum of two years experience teaching ESL 209 at the institution where the present study was conducted. For this selection task, these teachers were randomly divided into five groups of three. The three teachers in each group received the same set of five compositions representing one of the five possible letter grades (A, B, C, D, and F). None of the teachers in each group knew the letter grade of the compositions they were grading, or that their set represented compositions of the same letter grade. The five sets of five compositions were selected from the archives of ESL exams. Each composition had been given an A, B, C , D, or F grade by teachers who taught ESL 209 during the Winter 2003 session, using the Trait Scale and band descriptors described above. The three teachers in each of the five groups were asked to read their set of compositions and to mark them using the same Trait Scale and band descriptors used by the Winter 2003 in-course teachers.

Following the grading of the compositions, the papers were selected as representing models of an A, B, C, D, and F papers. Selection of these Models was based on how consistent the three raters evaluating the papers were in assigning a grade to them, one of the five letter grades mentioned above. Thus, a composition that had been

given the same letter grade by at least two teachers in the same group was selected as the Model for that particular letter grade. The five Models represented the middle range of each of these letter grades, that is an A, B, C, etc., grade rather than A+, A-, B+, B-, etc. The purpose of this stage was to select the Models that would be used as a rating instrument together with the Trait Scale during the grading of the Target Compositions (see Distribution of Target Composition stage below).

2. Distribution of Practice Compositions

In this stage, each participant met with the researcher for a 20-minute session. During this session, participants were given the two Practice Compositions, the Trait Scale and the band descriptors. The participants were also given written instructions on how to use the Trait Scale, and they were allowed to ask questions to clarify any uncertainty about its use. The participants had three days to grade and return the two compositions, whereupon the researcher recorded any differences in the spread of scores and used this, along with information gathered from the Background Questionnaire, to divide the participants into two equal groups (Trait Scale group and Trait Scale + Models group). As mentioned earlier, care was taken that each group included raters that had similar range of scores for the two Practice Compositions.

3. Distribution of Target Compositions

The second stage involved distributing the sixteen Target Compositions and evaluation instruments. The compositions and rating instruments were distributed to participants, who met with the researcher individually. During the meeting, the

participants each received sixteen Target Compositions but those designated to the Trait Scale group received one rating instrument, the Trait Scale and the descriptors only, to use in evaluating the compositions. Those who were designated to Trait Scale + Models group received the sixteen Target Compositions and two rating instruments—that is, the Trait Scale, descriptors and the Models. In the meetings for those in the Trait Scale group, the researcher led a review and discussion of the use of the Trait Scale and band descriptors. For those in the Trait Scale + Models group, the researcher reviewed and discussed the Models, and indicated how they were to be used in addition to the Trait Scale and band descriptors. Since the raters were already familiar with the Trait Scale, there was no discussion on how to use this testing instrument. Before the raters in the Trait Scale + Models group began grading the Target Compositions, they read all five Models and answered the *Pre-grading Questionnaire*.

Following the review and discussion of the testing instruments, the researcher asked the participant to grade the first composition in the test package (i.e., TC 1). The purpose of having the participant grade the first composition in front of the researcher was to ensure that the raters in the Trait Scale + Models group used the Models correctly since it was their first exposure to this testing instrument. For the sake of consistency, participants in the Trait Scale group were also asked to grade the first composition in front of the researcher. However, practice using the Trait Scale was not a concern for the participants in this group since they already had experienced using this testing instrument in courses they had previously taught in the department as well as during the distribution of the Practice Compositions stage. Three participants in the Trait Scale group (TS 1, TS 2 and TS 7) were not able to grade TC 1 during the discussion session because of time

constraints. All participants in the Trait Scale + Models group (the group of principal interest here), though, graded the first composition in front of the researcher.

The discussion session for all participants in both groups lasted one hour. At the end of the discussion session, the participants were instructed to grade the remaining set of compositions in the order in which they appeared (i.e., they were to grade TC 2, followed by TC 3, etc.) and to answer the debriefing questionnaire following completion of the grading of all sixteen Target Compositions. The participants were also instructed to return the graded compositions within one week. The teachers' ratings at this stage provided the comparison data that were used to answer the research questions about rater reliability and severity.

Chapter 4: Results

In this chapter, discussion first turns to the scoring procedure used in the study. This is followed with a report on the data analyses used to answer the first question on rater reliability. Finally, findings from the MFRM model and an answer to the question concerning the spread of rater severity are provided.

Scoring Procedure

As outlined in the Methodology section above, the Trait Scale required the participants to give each composition a letter grade ranging from A+ to D-, to F for each of the five traits (e.g., content, grammar, etc.) contained in the analytic grid. As the grid had differently weighted components (e.g., content is worth 15% of the total score, grammar is worth 50% of the total score, etc.), each letter was assigned a numeric value that corresponded to the weight that each component was accorded in the overall evaluation scheme. For example, as content is worth 15% of the total score, a score of A+ on this trait is, therefore, worth the full 15% and was assigned the numeric equivalent of 15. A rater who assigned the letter grade B in content also selected the corresponding numeric value 12 because a B in content is worth 80% of the trait and 12% of the total grade. A total score for each composition is the sum of the numeric values equivalent to the letter grades given for each of the five traits. For example, if a rater assigned an A+ to all the traits on the analytic grid, the corresponding numeric values (e.g., 15 for content, 15 for organization, 50 for grammar and language use, 15 for vocabulary, and 5 for mechanics) would be added to produce a total possible score of 100 for the composition.

The scores on the Trait Scale were subjected to statistical analyses and the results from these analyses were used for comparing the two groups of raters. For the group of raters that used the Trait Scale alone, their only set of scores used for analysis was the data obtained from the Trait Scale. For the group of raters who used the Trait Scale + Models, it was assumed that any effect of the Models on their evaluation of the target compositions would be manifested, if at all, in their scores indicated on the Trait Scale.

Pre-grading and Grading with Models Questionnaires

To ensure that the Trait Scale + Models group read, understood and used the Models to grade the Target Compositions, the raters in this group had to read all five Models and complete the *Pre-grading Questionnaire* before they began grading the Target Compositions. After reading each Target Composition and before grading each composition with the Trait Scale, the raters in this group completed the *Grading with Models Questionnaire* (See Appendix L).

The *Pre-grading Questionnaire* contained multiple choice questions, with the distractors carefully worded so that the raters had to read the composition in order to be able to select the correct responses. All the raters in the Trait Scale + Models group completed this questionnaire and for the most part answered the questions correctly. The mean score for the group was 81%. With respect to the responses on the *Grading with Models Questionnaire*, there was an overall agreement that each Target Composition was similar to the Model they selected as a grading guide. For example, the raters selected the values five, six or seven on a seven-point scale (where one has the value of "not at all representative" and seven "perfectly representative") 75% of the time, and this was true

for each of the traits (e.g., content, organization, etc.). Where raters indicated most that the Model selected did not represent the Target Composition they were grading was in content, with the values one, two or three selected 25% of the time.

Rater Reliability

The first research question investigated in this study was whether the use of Models and a Trait Scale brought about greater rater reliability than the use of a Trait Scale alone. To answer this question several analyses were conducted. These analyses involved computing a crosstabulation of the raw scores as well as calculating intraclass correlation coefficients for each group of raters. A *t*-test analysis was also conducted to determine significant differences, if any, between the two groups. Finally, the data was subjected to the Multi-Facet Rasch Measurement model to identify potential sources of variance that contribute to rater inconsistency.

Crosstabulation: The crosstabulation scores for each composition showed that 50% or more of the raters (i.e., at least 5 out of 10 raters per group) in the Trait Scale + Models group agreed on the same letter band (i.e., A range, B range, C range, D range or F range) 94% of the time, or for 15 of the 16 compositions. This was surprisingly high. The Trait Scale group showed 50% or more agreement on the same letter band 69% of the time, or for 11 of the 16 compositions. A crosstabulation of the total score was also conducted for the grades given along a 13-point scale—that is, for letter grades ranging from A+ to D-, or F. In this case, 50% or more agreement occurred in the Trait Scale + Models group in three of the 16 papers, and in Trait Scale group in one of the 16 papers. The results from these crosstabulation analyses indicate that the raters who used the Trait

Scale and Models to grade the 16 compositions had a tendency to agree more on their grades than raters who used the Trait Scale alone.

Intraclass correlations: Intraclass correlations were computed for each group of raters in order to determine rater reliability values. This procedure was done for three sets of test scores. The first set included total scores based on a 5-point scale—that is, grades based on the five letter bands, where, for the purpose of calculating the correlation coefficient, each letter was given a corresponding numerical value from 1 to 5 (e.g., F = 1, D = 2, etc.). The second set of scores were total scores based on a 13-point scale—that is, scores based on letter grades ranging from F, D- to A+. Again, for the purpose of computing the coefficient alpha for each group, these letter grades were given a corresponding numerical value (e.g., F = 1, D- = 2, D = 3, etc.). Finally, the correlation coefficients were calculated using total scores based on a 100-point scale—that is, the numerical values given out of a possible total score of 100.

Table 1 shows the intraclass correlation coefficients for the three types of total scores for the Trait Scale and Trait Scale + Models groups. For the 5-point scale, the single measure correlation for the Trait Scale group and the Trait Scale + Models group are .63 and .60, respectively. The correlation coefficients are slightly lower for both groups on scores from the 13-point scale where the α-value for the Trait Scale group is .55 and .62 for the Trait Scale + Models group. On the total scores from the 100-point scale, the coefficient alpha is .56 for the Trait Scale group and .64 for the Trait Scale + Models group. While the coefficients for the three types of scores for both groups were lower than expected, the reliability coefficient for the Trait Scale + Models group was still slightly higher than that of the Trait Scale group for the 13-point scale and raw

scores. The correlation coefficients were similar for scores taken from the 5-point scale because this type of scale would only reveal differences that show considerable variance. In other words, the more scores on the scale, the more variability among raters. The 100-point scale, then, has the greatest potential for variance, and, in fact, the difference found between the two groups was greatest when coefficient alphas were computed using scores on this scale.

Table 1—Intraclass Correlation Coefficients

Group	5-point Scale	13-point Scale	Raw Scores (Total 100 points)
Trait Scale	.63	.55	.55
Trait Scale + Models	.60	.62	.64

T-test: Analyses were conducted to determine whether the individual rater's mean score for the sixteen compositions he or she rated differed significantly from the mean score of their group. For example, a rater's mean total score (out of 100) was compared to the mean total score of the 10 raters in that group. This same procedure was also applied to the individual and group mean scores for each of the five traits (i.e., content, organization, grammar & language use, vocabulary and mechanics. *T*-tests were then conducted to compare each individual rater's total score per compostion with his or her group's mean overall rating for this same composition. Similar *t*-tests were also conducted comparing each individual rater's score for each of the five traits with his or her group's scores on these same traits.

Table 2—Results of *t*-test Analysis for Trait Scale Group

Asterisk (*) indicates *t*-test values that are significant.

Trait Scale Raters	Mean Total Score	Mean Content Score	Mean Organization Score	Mean Grammar & Language Use Score	Mean Vocabulary Score	Mean Mechanics Score
T1	70.9125	11.3750	11.9688	32.8750	10.4375	4.2563*
T2	63.3125*	9.6875	11.3750	34.8750	6.3750*	1.000*
T3	70.3688	10.5000	10.8750	34.4063	10.6563	3.9313
T4	73.0750	11.4375	11.5938	35.5000	10.7500	3.7937
T5	75.9125	12.1875*	12.4063*	35.3900	11.5938*	4.2500*
T6	66.2500*	10.6125	10.5438*	32.0438*	8.9688*	4.0813*
T7	77.2625	10.6875	11.3750	39.3750*	11.5313	4.2935*
T8	72.1875	10.5625	10.5938	35.2813	11.6563*	4.0938*
T9	78.9000*	12.1563*	12.3438	39.0000*	11.5313	3.8688
T10	72.4813	11.5188	11.6438	35.0687	10.6313	3.6188
Group Mean Score	72.0663	11.0725	11.4719	35.3900	10.4131	3.7188
df = 1, 15	* T< .002, Bonferroni					

Table 3—Results of *t*-test Analysis for Trait Scale + Models Group

Asterisk (*) indicates *t*-test values that are significant.

Trait Scale + Models Raters	Mean Total Score	Mean Content Score	Mean Organization Score	Mean Grammar & Language Use Score	Mean Vocabulary Score	Mean Mechanics Score
M11	74.3688	11.4375	11.2500	37.1563	10.9688	3.5563
M12	69.6938	10.6250	10.9375	34.2500	10.4063	3.4750
M13	67.4937	10.4375	10.5938	32.0000	10.4375	4.0250
M14	71.0000	10.0313	10.9375	36.3125	9.7500	3.9688
M15	69.1688	10.4063	11.3438	33.1875	10.4063	3.8250
M16	73.7750	12.7188*	13.1875*	33.3125	10.7188	3.8375
M17	69.7125	10.5313	10.8750	33.9375	10.7188	3.6500
M18	70.2375	11.3750	11.3438	34.0625	9.7188*	3.7375
M19	78.1750*	11.0938	12.4688*	38.8750*	11.9063*	3.8313
M20	72.2500	11.1563	10.8750	35.0000	11.1563	4.0625*
Group Mean Score	71.5875	10.9813	11.3813	34.8094	10.6187	3.7969
df=1, 15	* T< .002, Bonferroni					

The results of the *t*-test analysis for the Trait Scale group and the Trait Scale + Models group are presented above in Tables 2 and 3, respectively. The first column in each Table shows the ten raters of the group. The second, third, fourth, fifth, sixth and seventh columns present each rater's mean score on the six possible ratings given to the

compositions—that is, the overall score and the scores on content, organization, grammar & language use, vocabulary and mechanics. The second to last row in each column shows the group mean score for each of the six possible ratings. The bottom row presents the degrees of freedom (*df*) and the Bonferroni corrections for the alpha level, which was set at .002. An asterisk indicates those *t*-test values that are significant. With respect to Table 2, the number of Trait Scale group raters whose mean score significantly deviated from the mean score of their group ranged from two in content (T5 and T9) and organization (T5 and T6), to three in the total score (T2, T6 and T9) and grammar & language use (T6, T7 and T9), to four in vocabulary (T2, T5, T6 and T8), and finally to as many as six in mechanics (T1, T2, T5, T6, T7 and T8). In Table 3, on the other hand, the number of Trait Scale + Models group raters whose mean scores significantly deviated from the group mean ranged from only one in the total score (M19), content (M16), grammar & language use (M19) and mechanics (M20) to two in organization (M16 and M19) and vocabulary (M18 and M19). In addition, the Trait Scale + Models group raters whose scores deviated significantly tended to be the same individuals (M16 and M19). These results clearly show that there were more people in the Trait Scale group whose scores deviated from the group mean than in the Trait Scale + Models group. In other words, when comparing the scores of the raters using the Models to the scores of the raters using the Trait Scale alone, there was a tendency for the raters using the Models to be more consistent. Thus, in relation to the first research question, the *t*-test findings suggest that the use of pre-selected Models along with the Trait Scale may bring about a higher level of rater reliability than the use of the Trait Scale alone.

Rater Severity

The second research question is concerned with whether the use of Models and a Trait Scale reduces the spread of rater severity more than the use of a Trait Scale alone. In order to answer this question, results from the MFRM analysis were used. As already discussed above, raters can differ from one another in the severity with which they evaluate papers. Some raters, for example, can be more severe at grading so that they would distribute grades only from a narrow range in the trait scale or only from the bottom end of the trait scale. It was hypothesized that the use of Models and a Trait Scale would affect rater evaluations such that there would be a lesser spread of severity scores among those who used these model compositions compared to those who did not use them. In order to test this hypothesis, the two groups of raters' scores on each component of the Trait Scale for all 16 papers were subjected to the Minifac Facets computer program, a student version of FACETS (Linacre, 1989). Minifac Facets is used to execute the MFRM model.

Entering these scores for this analysis was achieved first by converting the letter grades for each dimension from the Trait Scale into numbers (or *numerical counts*). Since there were thirteen possible letter grades, there were thirteen possible numerical counts, where A+ at the top of the scale was assigned the numerical count 13 and F at the bottom of the scale was assigned the numerical count 1. Therefore, for the purpose of the Rasch analysis, all performance dimensions were equally weighted. The Minifac Facets software transformed the raw, numerical counts into probability measures by taking into account aspects (or *facets*) of the test setting that affect score outcomes. Though the MFRM model allows for the analysis of many facets, studies investigating rater severity

normally consider only three. These are candidate (i.e., composition) ability, rater severity, and item (i.e., trait) difficulty. The mathematical model used to calibrate the raw counts takes the following form:

$$\text{Log } [P_{nijk}/P_{nij(k-1)}] = B_n - D_i - C_j - F_k$$

where

P_{nijk} = probability of ratee n being rated k on item i by rater j

$P_{nij(k-1)}$ = probability of candidate n being rated k – 1 on item i by rater j

B_n = ability of candidate n

D_i = difficulty of item i

C_j = severity of rater j

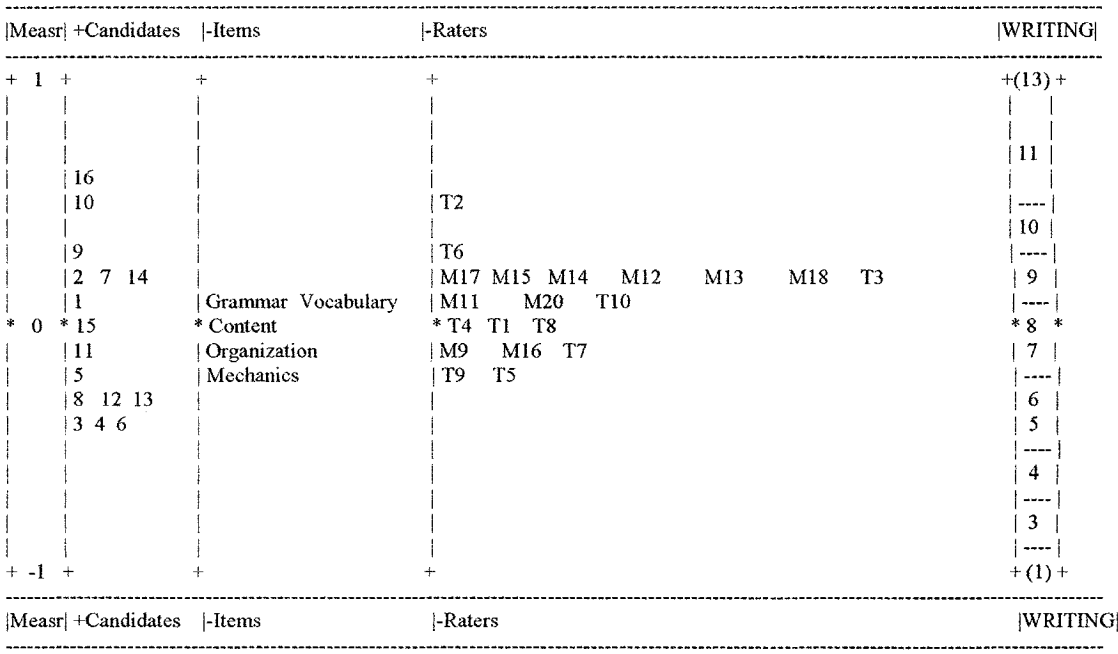
F_k = difficulty of scale category k relative to scale category k – 1.

In addition to calibrating the raw counts, Minifac Facets plots the estimates onto a linear, equal interval *logit* scale. Figure 1 below shows an example of an equal interval logit scale with plotted measures for three facets: person ability, item (i.e., trait) difficulty and rater severity. Column 1 is the logit scale, which is the same for all facets in the data set. Column 2 shows the 16 candidates (i.e., compositions in this case) that were graded, and these are ordered from most "able" at the top of the scale to least able at the bottom.

To find out whether differences existed in the spread of rater severity between the two groups one has first to look more carefully at Column two in Figure 1. This column shows, for example, that composition 16 was scored highest. Descriptive statistical tests performed on the data show 70% of the raters in the Trait Scale group and 90% of the raters in the Trait Scale + Models group gave this composition a grade in the A letter band (i.e., A-, A, or A+). The third column shows the five trait items (e.g., contents, organization) ordered from most severely rated at the top of the scale to least severely

rated at the bottom of the scale. In this case, grammar and language use and vocabulary were graded most severely by all twenty raters, while mechanics was graded least severely.

Figure 1—Logit Scale Showing Candidate Ability, Item Difficulty and Rater Severity Measures



The fourth column shows the raters who are ordered in terms of the most severe rater (T2) at the top of the scale and the least severe, in this case two raters (T5 and T9), at the bottom. From this column, we can see that the Trait Scale raters are more spread out on the scale, with some found at the extreme ends scale. In other words, those who are identified as most severe and least severe in their ratings are raters from the Trait Scale group. The Trait Scale + Models group, in contrast, tend to be more clustered together, with only two (M16 and M19) deviating from the group. It is interesting to note

that these two deviant raters are the same two raters whose mean scores significantly deviated from their group's composite mean most often, as earlier identified in the *t*-test analysis. Furthermore, the Trait Scale raters who significantly deviated from their group mean most often (i.e., T2, T6, T5 and T9) are also the same ones identified to be at the extreme ends of the logit scale in this analysis. This finding means that the results of the *t*-test conducted to find out which raters deviated from their group are supported by the results of the MFRM analysis and confirms that there are more Trait Scale group raters deviating from the group mean for each composition than Trait Scale + Models group raters.

Table 4—Raters' Mearsurement Report (Trait Scale Group)

RATERS	MEASURE (logits)	MODEL STANDARD ERROR	INFIT MEAN-SQUARE
T2	.47	.05	2.29
T6	.25	.04	.92
T3	.14	.04	.61
T10	.08	.04	1.00
T4	.04	.04	.65
T8	.03	.04	.91
T1	-.01	.04	.60
T7	-.08	.04	1.23
T9	-.12	.05	1.03
T5	-.20	.05	.76
RMSE (Model) = .04 Adj. S.D. = .18 Separation = 4.05 Reliability = .94 Standard Deviation (Measure) = .18			

In considering the two groups separately, Table 4 above shows the Raters' Measurement Report, which provides a more detailed Rasch analysis of rater behaviour, for the Trait Scale group. The first column presents the teachers who are ordered from

most severe at the top to least severe at the bottom. The second column provides the logit measures for each rater. Again, these measures are derived from a logarithmic equation and report the probabilities of a particular response. In this case, the logit measures indicate the extent to which raters are severe or lenient in their judgements. Those raters who have higher logit measures are more severe overall, while those who have lower values are more lenient. Therefore, T2 is the most severe in the group at .47 logits and rater T5 is the least severe at -.20 logits. The third column gives the model standard error, which is fairly low given the number of data points and the fact that there is no missing data in the data set. The fourth column provides the infit mean-square statistic for each rater, and will be discussed further below.

Referring once again to column two, which provides the logit measures, a spread in severity for the raters in the Trait Scale group ranges from .47 to -.20, or a difference of .67 logits. This spread in severity is smaller for the Trait Scale + Models group, as Table 4 shows a spread in severity from .28 to -.18, or a difference of .46 logits. This difference in the range of severities is also reflected in the standard deviations and separation indices of the two groups. At the bottom of Tables 4 and 5 are found standard deviations and the separation indices for their respective groups of raters. The separation index is the ratio of the corrected standard deviation (Adj. S.D.) of element measures (in this case, raters) to the root mean-square standard error (RMSE), and it indicates variance among raters. A value of equal to or less than 1 would indicate that the raters were equally severe in their ratings because the adjusted standard deviation should be equal to or smaller than the root mean-square standard error (RMSE) of the entire data set. The value of 4.05, on the other hand, indicates variance among Trait Scale group raters was

slightly over four times the error of estimates. This variance was less in the Models + Trait Scale group, where Table 5 shows variance at slightly less than three times the error of estimates. Comparing the two groups with respect to their standard deviations, we see that in Tables 4 and 5, respectively, the Trait Scale group raters are slightly higher at .18, compared with .15 for the Trait Scale + Models group. One final statistic that can be found at the bottom of Tables 4 and 5 is the reliability index. This indicates the degree to which the analysis reliably separates elements within a facet, which in this case are the different levels of severity among raters in the Trait Scale and Trait Scale + Models groups. A value closer to 0 would indicate that the raters are behaving equally in severity and a value closer to 1 would indicate substantial differences in scores assigned to the compositions, which is indeed the case for the Trait Scale and Trait Scale + Models groups at .94 and .89, respectively.

Table 5—Raters' Mearsurement Report (Trait Scale + Models group)

RATERS	MEASURE (logits)	MODEL STANDARD ERROR	INFIT MEAN-SQUARE
M12	.28	.05	.86
M17	.25	.05	.91
M13	.24	.05	.74
M18	.22	.05	.87
M14	.21	.05	.1.14
M15	.20	.05	.82
M20	.07	.05	.88
M11	.06	.05	1.24
M16	-.12	.05	1.46
M19	-.18	.05	1.03
RMSE (Model) = .05 Adj. S.D. = .14 Separation = 2.80 Reliability = .89 Standard Deviation (Measure) = .15			

Referring, once again, to the fourth column of Tables 4 and 5, the infit mean-square statistic is a measure of how consistent a rater is within himself—that is, it is a measure of intra-rater consistency. Rasch analysis calculates this value by comparing predicted, or *expected*, scores to the raw, or *observed*, values. Before calculating fit statistics, though, the analysis calibrates predicted scores, which is achieved by taking into account the severity of each rater, the difficulty of each item and the ability of each candidate. Through successive iterations it compares expected and observed responses and refines these estimates until there is a sufficiently close match between the expected and observed scores. The fit statistics, then, summarize for each element within a facet, in this case the raters, the extent of fit between expected and observed values. In the case of the facet for raters, these statistics show the fit between expected and observed ratings.

Though Rasch analysis provides several fit statistics, the infit mean-square residual is considered most informative by researchers (McNamara, 1996). Since these residuals have an expected value of 1, individual values above 1 indicate greater variation than expected, and values below 1 indicate less variation than expected. Although there is no fixed standard for determining the degree of fit, McNamara (1996) identifies the fairly conservative range between .75 and 1.3 as a standard to apply. Therefore, any individual value greater than the set limit indicates greater variation than expected and is identified as a *misfitting* item. A value less than the limit set indicates less variation than expected and is identified as an *overfitting* item.

Applying the above standards to the results of this study one can see that a rater (T2) in the Trait Scale group was misfitting, or behaving significantly unpredictably in his grading while three other raters (T3, T4 and T1) were overfitting, or behaving

significantly overly predictably in their grading. Though the overfit statistic may sound like these raters are exceptionally accurate in their judgements, the more likely reason is that they are underusing, or overusing, steps (i.e., grade bands) in the rating scale (McNamara, 1996). To confirm this, one needs to refer to the raw data and see that the pattern of grading of the three overfitting raters is more deterministic than probabilistic. For example, T1 avoided using the bottom of the thirteen-point scale for content, organization and mechanics, and avoided using the top end of the scale when grading grammar and language use and vocabulary. In other words, T1 did not give the grades F, D-, D or D+ for content, organization and mechanics nor the grades A-, A and A+ for grammar and language use and vocabulary on any of the 16 compositions. Similarly, T3 avoided using the upper end of the scale—that is, this rater did not give a letter grade in the A band—on any of the 16 compositions for all traits with the exception of mechanics. T4, to a similar but lesser extent than T1 and T3, showed a tendency to avoid giving extreme grades (i.e., F, D-, A and A+) and showed a preference for giving grades in the C and B letter bands. In other words, by reviewing the raw scores of the three raters who were identified by the MFRM model as overfitting, a pattern emerges where there is a preference to give grades that cluster around the middle of the grading scale despite a range of compositions of different "ability" levels being graded.

This is particularly revealing when one looks at the overall use of the rating scale by the ten Trait Scale group raters. Figure 2 below summarizes the overall use of the rating scale by both groups and illustrates that the Trait Scale raters failed to discriminate the C and B letter bands—more specifically, they avoided discriminating between the letter grades C, C+, B- and B. In fact, the Trait Scale group raters were only

discriminating seven steps of the thirteen-step rating scale used to grade the compositions. The overfitting statistic provided by the MFRM model, then, highlights those raters who are contributing most to this group tendency. The model also indicates that a source of measurement error leading to rater inconsistency lies in the central tendencies—that is, the overuse of the middle steps on the rating scale—by these three individual raters.

Figure 2 — Use of 13-point Rating Scale by the Trait Scale Group and Trait Scale + Models Group Raters on 16 Compositions

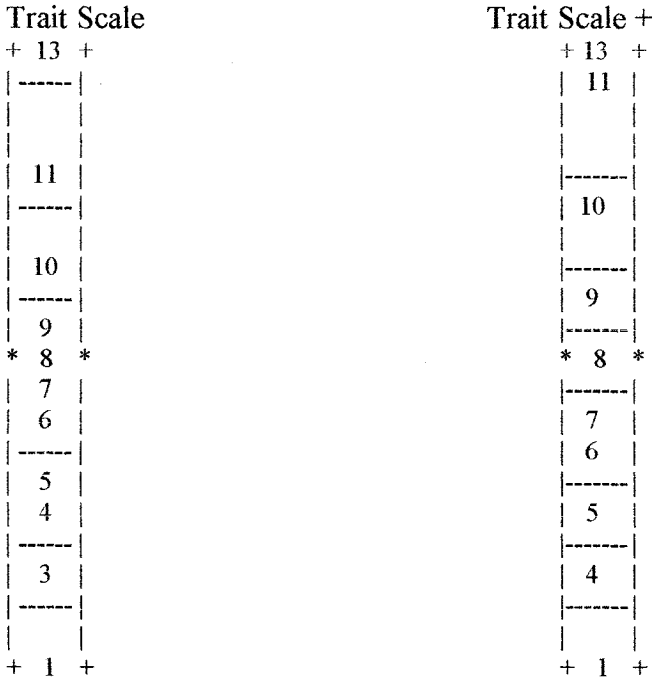


Table 5 provides statistics for the Raters' Measurement Report for the Trait Scale + Models group. In this case, raters range in severity from M12, the most severe rater in the group at .28 logits, to M19, the most lenient rater at -.18 logits. The table shows the standard error at .05 to be low, suggesting that the calibrated estimates are accurate.

However, unlike the Trait Scale group which had four raters who were identified as either misfitting or overfitting, the Trait Scale + Models group shows only one rater who misfitted (M16) and another who was a borderline overfit (M13). Examination of the raw data shows that M13 avoided giving grades in the A letter band for four of the five traits on almost all sixteen compositions, composition 16 was the only paper to which M13 gave an A grade. The finding that fewer raters in the Trait Scale + Models group are identified as misfitting or overfitting compared to the Trait Scale group suggests that the raters in the former group discriminated steps in the grading scale more often than those in the latter.

Figure 2 also shows that the Models + Trait Scale group raters used one more step in the rating scale (i.e., eight steps) than the Trait Scale group raters. More importantly, they discriminated between the C and B letter grades better than the raters in the Trait Scale group. With respect to the separation index, the Trait Scale + Models group is slightly under three times the error of estimates at 2.80, and the reliability index is also slightly lower than that of the Trait Scale group at .89. The MFRM findings support the *t*-test results that indicate the Models + Trait Scale group is behaving more consistently than Trait Scale group as well as provide further insight into rater behaviour contributing to score variance.

To summarize the results of the analysis designed to find an answer to the first research question on rater reliability, the *t*-test identified more significant score variance among the raters who used the Trait Scale alone than those who used Models in addition to the Trait Scale. This result was confirmed by the MFRM analysis, which identified the same raters deviating the most from their respective groups. Furthermore, the MFRM

analysis identified the Trait Scale group raters as those who were most severe and most lenient in their ratings—that is, those who were at the extreme ends of the logit scale of the set of twenty raters. The Rasch model also showed that none of the raters graded the sixteen compositions equally severely. The fit statistics, also provided by the Rasch model, identified a central tendency that was more pronounced among Trait Scale group raters. To answer the second question on the spread of rater severity, the separation indices showed that there is greater variance among raters in the Trait Scale group than those in the Trait Scale + Models group, and that the difference in this variation is more than a full point. In relation to the second question, then, these statistics indicate that the use of Models to grade student essays reduces the spread of rater severity. In short, while the raters in both groups are not behaving consistently, there is a tendency for the raters in the Trait Scale + Models group to vary less in their judgements than those in the Trait Scale group.

Chapter 5: Discussion

In this chapter, a summary of the findings is presented. This will be followed by remarks on the contributions this study makes to research investigating rater reliability, rater severity and rater consistency. Finally, limitations and future directions will be noted.

Summary of findings

Cross tabulations of the total scores showed that the raters who used the Models were more often in agreement with each other than the group of participants who scored the compositions using the Trait Scale alone. The results of intraclass correlation analyses also showed slightly higher reliability coefficients for the group of raters who used the Trait Scale + Models than for the group that used the Trait Scale alone. In addition, *t*-test analyses compared each rater's mean rating for the 16 compositions he or she was assigned against the mean rating of the entire group. Results from these analyses showed a tendency for more of the raters' scores in the Trait Scale group to deviate significantly from their group mean score more often than was the case with the raters' scores in the Trait Scale + Models group. Moreover, Trait Scale + Models raters whose scores deviated from the group mean tended to occur only for two particular raters. Overall, there were fewer deviant scores in the Trait Scale + Models group.

An MFRM (Rasch) analysis supported the *t*-test findings by identifying these same two teachers as deviating from the rest of the group in terms of their overall severity measures. Furthermore, when comparing the overall severity measures for the twenty raters, those identified as grading the sixteen compositions either most severely or least

severely were raters from the Trait Scale group. The Trait Scale + Models group rater severity measures, on the other hand, were more clustered together, even when taking into account those two raters who deviated from their group. This suggests that the Models had a positive effect on the spread of rater severity, The Rasch analysis also revealed other differences between the two groups. First, the MFRM analysis showed that there were more Trait Scale raters who were deterministic, or too predictable, in their grading. Second, the Trait Scale + Models group was better at discriminating steps of the rating scale than the Trait Scale group. All these findings suggest that the Models helped raters grade compositions more consistently than if they were to use a Trait Scale alone.

Rater reliability

One purpose of this study was to investigate the extent to which two testing instruments (i.e., models and rating scale) affect rater reliability in a direct ESL writing assessment context. Both testing instruments represent ways to introduce a set of object standards for the purpose of controlling rater variance. Given that creating objectified standards is a necessary condition for attaining high levels of rater consistency (Gamaroff, 2000), it is important to determine to what extent these attempts at establishing such criteria can improve the consistency of raters' scores. One way to interpret rater consistency is see it as rater reliability—that is, to understand sources of measurement errorS that contribute to variance as homogenous and random. Findings in the literature have suggested that analytic scales bring about higher levels of rater reliability than holistic scales (e.g., Bacha, 2001). This study contributes to the literature

by suggesting that an analytic trait scale used with model compositions can make raters even more consistent than if they were to use the analytic scale alone.

This finding is particularly important because so far no direct ESL writing assessment study has considered to what extent rater variance is affected by the use of model compositions and an analytic rating scale. While research in other contexts has suggested that models can be effective in controlling rater variance (e.g. Dandenault, 1997), their usefulness in direct ESL writing assessment seems to have been assumed. Based on the finding that the group that used the Trait Scale + Models had a higher coefficient alpha than the group that used the Trait Scale alone suggests that model compositions may have helped raters interpret the evaluation criteria more consistently. The models may have provided the raters with clear examples and helped clarify any misconception as to the difference between a B paper and a C paper, and so on. The models may also have helped raters interpret more precisely the evaluation criteria found on the rating scale because the models contain specific examples to illustrate those characteristics of written language used in various grades of ability on which raters based their judgments.

Still, the coefficient alpha for the Trait Scale + Models group at .64 is below the commonly accepted value of .80 (Shohamy et al. 1992; Engelhard, 1992; Bacha, 2001). This, however, does not diminish the potential for using model compositions. Studies that have reported satisfactory levels of rater reliability expressed as a correlation coefficient have either applied or assumed intensive procedural training. For example, the raters in Shohamy et al.'s (1992) study underwent a training session that consisted of lengthy discussions of each evaluated sample and negotiation until consensus was

reached. Rater training is important because it brings raters into agreement on established objective standards and satisfies another condition that is necessary if high levels of rater consistency are to be attained (Gamaroff, 2000). Based on Bachman & Palmer's (1996) guideline for general rater training procedures, models represent pre-rated language samples that are used along with the rating scale. During this training both testing instruments are read, applied and discussed. The raters who participated in the present study were not exposed to a training session. It seems very likely, then, that using model compositions along with an analytic scale in a rater training procedure should further improve rater reliability.

Rater severity

A second purpose to this study was to determine how the models and the rating scale influence the spread of rater severity. Studies investigating rater consistency from this perspective have not considered the effect of model compositions on rater severity, nor have they looked at the usefulness of rating scales in this respect. Rather, these studies have focused on the effects of rater training, or have assumed procedural rater training prior to the study. These studies have consistently reported that rater severity differences survive training (Wigglesworth, 1993; Lumley and McNamara, 1995; McQueen and Congdon, 1997; Weigle, 1998). Weigle (1998), nonetheless, reported a reduced spread in rater severity differences after training. Still, the primary goal of these studies has been to focus on the finding that rater severity differences are part of the direct assessment process and cannot be eliminated. Furthermore, some proponents of the Rasch measurement model hold the view that the most effective way to deal with

inter-rater severity differences is to allow software, like Minifac Facets, to compensate for the inequalities and balance scores so that raters' assessments are fairer to the students (McNamara, 1996). This study, unlike the similar rater severity studies cited above, set out to determine differences in the spread of severity measures when comparing two testing instruments commonly used in rater training procedures. It may very well be the case that severity measures across raters will never be equal. Still, focussing efforts on reducing this inter-rater severity spread reduces score variance and improves rater consistency. The findings of this study reveal that there is a smaller overall severity spread across raters when compositions are graded with the Trait Scale + Models than with the Trait Scale alone. This is a particularly important finding when considering that one of the criteria for equally dividing the two groups of raters was to take into account the spread of the score differences observed in the grading of the two Practice Compositions. In other words, raters who graded the Practice Compositions similarly were equally divided into two groups. The reason that the Models used along with the Trait Scale may have reduced the spread of overall rater severity measures, once again, rests on the fact that the Models helped raters interpret the scoring criteria because they represented authentic examples of the target compositions.

Understanding rater consistency

The explanation for understanding rater consistency as rater reliability and rater severity is found in the statistical procedures researchers adopt to make inferences from their data sets. The present study applied traditional statistical tools (i.e., intraclass correlation and *t*-tests), which are computed using raw test scores, and a more recent tool,

the MFRM model, which calibrates the raw scores into estimates by taking into account candidate ability, item difficulty and rater severity, and plots these measures onto a true interval scale. Past studies in the field of large-scale direct writing assessment have cautioned against relying solely on satisfactory levels of rater reliability represented in the form of correlation coefficients (e.g., Engelhard, 1992). While interrater correlations can be high, differences in the spread of rater severity, for example, may be such that test takers are not always receiving grades that reflect their true ability. The result is that some test takers may fail while others of equal ability pass.

In addition to giving overall severity measures, the MFRM model provides statistics that offer further insight into rater behaviour contributing to score variance. For example, the fit statistics identified the Trait Scale + Models group raters as discriminating steps of the rating scale better than the Trait Scale group. This information is particularly important for high-stakes tests, such as the final exams graded at the institution where the present study took place. Graduate students enrolled in the ESL courses require the letter grade of B+ to pass. The raters in the Trait Scale group did not discriminate between the C and B letter bands; yet raters in the Trait Scale + Models group did. This suggest that the use of Models in addition to the Trait Scale improves raters' discrimination of the bands on the rating scale. Therefore, the findings of this study also contribute to the literature by suggesting that model compositions help raters discriminate more steps of a rating scale. Again, a reason for this lies in the fact that the model compositions represent examples of the five letter bands, thus helping the raters interpret the steps of the rating scale more consistently.

Without the MFRM analytical tool, such rater behaviour would not have been so clearly evident. Indeed, as proponents of the Rasch measurement model point out, basing inferences of rater consistency from analyses using classical true score models has its limitations (Linacre, 1989; McNamara, 1996). Moreover, the diagnostic information provided by the Rasch model has benefits for future research. This study has shown that both statistical procedures—that is, traditional tools and the MFRM model—can in fact be complementary.

Another strength of this study is that it offers a new perspective from which rater behaviour can be studied. There is no doubt that rater consistency in a direct writing assessment context is a complex and elusive field of study. For one, there exist numerous variables that can influence rater behaviour. From features of writing (e.g., grammar, vocabulary, etc.), to raters' background, to elements of the test situation, all of these factors can potentially influence score variance. In addition, some studies have suggested that rater consistency may in part be affected by the way in which rating scales are developed (Turner & Upshur, 2002). Furthermore, if the scale does not cover all the eventualities found in the test samples (e.g., clarifying grammatical and lexical nuances), as Lumley (2002) warns, raters are left to making decisions where they must reconcile "the rules" and their "intuitive impressions". Indeed, judging the more direct forms of writing performance in a communicative language framework involves complex mental processes because raters are required to weigh a multitude of variables. It makes sense, then, that providing raters with authentic compositions (i.e., models) that contain those variables upon which raters base their judgements reduces reliance on their intuitive impressions. Furthermore, the fact that model compositions represent the various grades

of assessment criteria (e.g., A grade, B grade, etc.) helps raters discriminate language ability. In short, the effect of model compositions on rater consistency is a new approach from which researchers can investigate ways to reduce score variance in a direct ESL writing assessment context.

Limitations

There are several limitations to this study. First, this study did not implement a pre-post design with a third group of raters that graded with Models only. Studies implementing a pre-post design could provide findings that showed a stronger causal relationship between the use of model compositions and rater consistency if the raters using the Models were found to have more consistent scores in the post design. Furthermore, using three groups could indicate whether or not using models in addition to an analytic scale is more useful than using models alone. Using a third group that graded with models alone beckons the question of how they should grade these compositions. Future studies could have raters grade the compositions analytically—that is, the raters would give a grade to each of the five components as those found on the Trait Scale. A second limitation of the study relates to the number of participants. Twenty raters (ten per group) were used because this was the minimum number that would allow for the findings to carry any weight. Using more than twenty participants would strengthen findings for the correlation and *t*-test analyses.

A third limitation is that this study did not consider whether Models have an effect on experienced versus inexperienced raters. The studies of Shohamy et al. (1992) and Weigle (1994) reported that training had a positive effect on score variance and that this

was particularly true for inexperienced raters. Moreover, both studies found no clear distinctions in score variance between experienced and inexperienced raters following training. While the present study used experienced raters—that is, raters who were familiar with the Trait Scale—it would be interesting to know the degree to which model compositions influenced rater consistency for inexperienced raters, and whether differences in score variance existed between experienced and inexperienced raters where both groups used models compositions. Furthermore, Weigle (1998) found that inexperienced raters gave more extreme scores and applied the rating scale more severely before training. Thus, rater training improved rater severity measures for inexperienced raters. It would be interesting to know whether model compositions had a similar effect on the severity measures of inexperienced raters.

Future directions

Again, future studies that implemented a pre-post design with a third group of raters and added more participants to each group could provide stronger findings. It would be beneficial if future studies investigated the effects of model compositions that represented all bands of the rating scale (i.e., in this case, the thirteen steps of the Trait Scale: A+ to D- and F). Again, the Models used in this study represented the five letter grades of A to D and F. Studies that use thirteen models (corresponding to each of the 13 grade bands), each representing a step in the rating scale, could indicate whether models that represented all possible bands of the scoring rubric improved rater reliability expressed as a correlation coefficient. These studies could also show whether these models improve rater reliability to a satisfactory coefficient value—that is, to a value

models improve rater reliability to a satisfactory coefficient value—that is, to a value where the correlation coefficient equalled or was greater than .80. With respect to rater severity, and the application of the Rasch model, investigating the effects of models that represented all levels of the scale could indicate whether these models further reduced rater severity spread differences, and whether they could further improve raters' discrimination of the steps on the rating scale. Of course, having an experienced group of raters come to consensus on thirteen model compositions may prove to be a challenging task.

Finally, qualitative research in the form of think-aloud protocols (e.g., Weigle, 1998), would support any quantitative findings as well as point to possible reasons why raters are more consistent when grading with model compositions. Such protocols could indicate which elements found in the models, as well as the test samples, rating scale and band descriptors, raters focus on when judging written performance. Having access to this information could indicate how model compositions help reduce score variance. It could also indicate which testing instruments raters focus on more when making their judgments—that is, we could discover whether raters base their decisions more on the rating scale or on the models when using both instruments to evaluate essays. We could also determine why and under what conditions either case is true. All this information gathered from qualitative analyses could, if needed, lead to improved testing instruments as well more effective procedures for their use in rater training sessions.

Rater training sessions can be time consuming and costly. Administrations could benefit by knowing whether the use of model compositions that accurately represented all bands of a rating scale led to satisfactory coefficient values and to what extent these

models reduced rater severity spread differences. Finally, the study indicates that administrations can benefit from the use of the MFRM model. The identification of individual raters who may be inconsistent, or too predictable, in their judgements would lessen costs by providing administrators with the diagnostic information necessary to determine those teachers that require rater training as well as identify those problematic areas on which training procedures should be focused.

Indeed, the factors influencing rater consistency are complex and attempts to attain the objective precision these direct, high-stakes tests demand have been elusive. This is why more research in the field of direct ESL writing assessment is needed if raters' inherent subjective judgements are to be reconciled. Conducting such research will not only inform theory but provide practical and feasible solutions to administrators and at the same time lead to fairer tests for our students.

Endnotes

[1] For a description of classical 'true' score measurement theory, see Bachman (1990), pp. 167-187.

[2] For review of Rasch measurement models, see Wright and Mok (2004).

References

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System 29*, 371-383.

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

Bachman, L.F. and Palmer, A.S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics 1*, 1-47.

Cumming, A. (1989). Writing expertise and second language composition. *Language Learning 39*, 81-141.

Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing 7*, 31-51.

Dandenault, E.J. (1997). Self-assessment of communicative ability: Investigation of a novel tool for ESL learners. Unpublished M.A. thesis. Concordia University.

Engelhard, G. Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education 5*(3), 171-191.

Gamaroff, R. (2000). Rater reliability in language testing: The bug of all bears. *System 28*(1), 31-53.

Hamp-Lyons, L. (1991). The writer's knowledge and our knowledge of the writer. In L. Hamp-Lyons (ed.), *Assessing second language writing in academic contexts* (pp. 15-36). Norwood, NJ: Ablex.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication 41*(2), 201-213.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing 19*(1), 3-31.

Linacre, J.M. (1989). *Multi-faceted measurement*. Chicago, IL: MESA Press.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing 19*(3), 246-276.

Lumley, T., Lynch, B.K. and McNamara, T.F. (1994). A new approach to standard setting in language assessment. *Language Testing 3*(2), 19-40.

Lumley, T. and McNamara, T.F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*(1), 54-71.

Lunz, M.E. and Stahl, J.A. (1990). Judge consistency and severity across grading periods. *Evaluation of the Health Professions 13*(4), 425-444.

McNamara, T.F. (1996). *Measuring Second Language Performance*. London: Longman.

McNamara, T.F. and Adams, M.J. 1991/1994: Exploring rater behaviour with Rasch techniques. Paper presented at the 13th Language Testing Research Colloquium, Educational Testing Service, Princeton, NJ, 21-23 March. (ERIC Document Reproduction Service No. 345 498)

McQueen, J. and Congdon, P.J. (1997). Rater severity in large-scale assessment: Is it invariant? Paper presented at the Annual Meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. 411 303)

Moss, P. (1994). Can there be validity without reliability? *Educational Researchers 23*(2), 5-12.

Shohamy, E., Gordon, C. and Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal 76*, 27-33.

Spaan, M. (1993). The effect of prompt on essay examinations. In D. Douglas and C. Chapelle (eds.), *A new decade of language testing research* (pp. 98-122). Alexandria, VA: TESOL.

Sweedler-Brown, C.O. (1993). ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing 2*(1), 3-17.

Tedick, D. and Mathison, M. (1995). Holistic scoring in ESL writing assessment: What does an analysis of rhetorical features reveal? In D. Belcher and G. Braine (eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 205-230). Norwood, NJ: Ablex.

Turner, C.E. and Upshur, J.A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly 36*(1), 49-79.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (ed) *Assessing Second Language Writing in Academic Contexts* (pp. 111-26). Norwood, NJ: Ablex.

Weir, C.J. (1990). *Communicative language testing*. NJ: Prentice Hall Regents.

Weigle, S.C. (1994). Effects of training on raters of ESL compositions. *Language Testing 11*(2), 197-223.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing 15*(2), 263-287.

Weigle, S.C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing 6*(2), 145-178.

Weigle, S.C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing 10*, 305-35.

Wright, B.D. and Mok, M.M.C. (2004). An overview of the family of Rasch measurement models. In E.V. Smith and E.M. Smith (eds.), *Introduction to Rasch measurement*.(pp. 1-25). Maple Grove, MN: JAM Press.

APPENDIX A

PARTICIPANT BACKGROUND INFORMATION

First name: _____ Date of birth: _____

Last name: _____ Gender: M / F

Please write requested information in the space provided. If more space is needed, please use the reverse side of either page of this questionnaire.

A. Language Background

Please indicate <u>your</u> first language as well as any other languages <u>you</u> speak and/or write. Please print clearly.

First language: _____

Other languages

Spoken: _____

Written: _____

B. Educational background

I. Degrees completed. *Please use "X" to indicate your response.*

☐ TESL Certificate ☐ Bachelor of Arts ☐ Bachelor of Education

☐ M.A. Applied Linguistics ☐ Ph.D(field of study): _____

☐ Other (Please specify): _____

II. Teacher training experience. *Please check (X) where appropriate.*

☐ Elementary ☐ High school ☐ Cégep ☐ University

☐ Other (Please specify):_____

Please check (X) the type of class you practice-taught during your teacher training as well as the first language of the learners.

☐ Homogenous class(es)
First language of learners: _____

☐ Heterogeneous class(es)
More common first languages of learners: (Specify the first five or six, if appropriate)

_____

Proficiency level (s): _____

C. Teaching experience

Please check (X) the type of institution, and indicate the number of years of teaching experience as well as the class-type, first language and proficiency level of the language learners.

☐ Elementary
- ☐ less than 1 year
- ☐ between 1 and 3
- ☐ 3-5
- ☐ 6-10
- ☐ more than 10

☐ High school
- ☐ less than 1 year
- ☐ between 1 and 3
- ☐ 3-5
- ☐ 6-10
- ☐ more than 10

☐ Cégep
- ☐ less than 1 year
- ☐ between 1 and 3
- ☐ 3-5
- ☐ 6-10
- ☐ more than 10

☐ University
- ☐ less than 1 year
- ☐ between 1 and 3
- ☐ 3-5
- ☐ 6-10
- ☐ more than 10

☐ Other (Please specify): _____
- ☐ less than 1 year
- ☐ between 1 and 3
- ☐ 3-5
- ☐ 6-10
- ☐ more than 10

☐ Homogenous class (es)
First language of learners: _____

☐ Heterogeneous class (es)
More common first languages of learners: (Please name the first five, if appropriate)

_____

Proficiency level(s):_____

D. Experience teaching composition

Please check (X) the type of institution where you taught composition, and indicate the number of hours of experience as well as the first language and proficiency level of the learners

☐ High school
- ☐ less than 50 hours
- ☐ 50-100
- ☐ 100-200
- ☐ 200-500
- ☐ more than 500

☐ Cégep
- ☐ less than 50 hours
- ☐ 50-100
- ☐ 100-200
- ☐ 200-500
- ☐ more than 500

☐ University
- ☐ less than 50 hours
- ☐ 50-100
- ☐ 100-200
- ☐ 200-500
- ☐ more than 500

☐ Other (Please specify): _____
- ☐ less than 50
- ☐ 50-100
- ☐ 100-200
- ☐ 200-500
- ☐ more than 500

First language(s): _____

Proficiency level(s):_____

E. Experience using a rating scale to grade ESL compositions

Please check (X) the type of rating scale you PREDOMINANTLY used to grade ESL compositions.

Note: A holistic scale is one in which the grade for the composition is based on a single score. An analytic scale is one in which the grade is the sum of separate scores given to the different components of writing (e.g., CONTENT, GRAMMAR, etc.)

Institution	Holistic scale *(single score)*	Analytic scale *(sum of separate scores)*
High school		
Cégep		
University		
Other (Please specify): _____		

F. Concordia University teaching experience

Please check (X) the course(s) and number of terms taught (e.g., Fall, Winter, etc.)

☐ ESL 208
- ☐ 2 terms or less
- ☐ 3 to 4 terms
- ☐ 5 to 10
- ☐ 10 to 20
- ☐ more than 20

☐ ESL 209
- ☐ 2 terms or less
- ☐ 3 to 4 terms
- ☐ 5 to 10
- ☐ 10 to 20
- ☐ more than 20

APPENDIX B

Debriefing Questionnaire

Part 1: Focus on the Grid.
Please use "X" to check your response.

1A. Did you use the grid while evaluating the compositions? ☐ YES ☐ NO
B. Did you consult the band descriptors while evaluating the compositions? ☐ YES ☐ NO

2A. Which component of the grid was the easiest to evaluate? Please check one only.
☐ Content ☐ Organization ☐ Grammar & Language Use
☐ Vocabulary ☐ Mechanics ☐ None of these components

B. Which component of the grid was the most difficult to evaluate? Please check one only.
☐ Content ☐ Organization ☐ Grammar & Language Use
☐ Vocabulary ☐ Mechanics ☐ None of these components

3A. Did you find any of the band descriptors vague or incomplete, thereby causing you to waver in your judgement of a grade? ☐ YES ☐ NO

B. Please specify which element(s) of the band descriptors you found vague or incomplete.

_____

_____

_____

_____

_____

_____

_____

4. In addition to the grid and band descriptors, do you think it would have been useful to have had models of an A paper, B paper, etc. to grade the compositions? ☐ YES ☐ NO

APPENDIX C

Debriefing Questionnaire

Part 1: Focus on the Grid.
Please use "X" to check your response.

1A. Did you use the grid while evaluating the compositions? ☐ YES ☐ NO
B. Did you consult the band descriptors while evaluating the compositions? ☐ YES ☐ NO

2A. Which component of the grid was the easiest to evaluate? Please check one only.
☐ Content ☐ Organization ☐ Grammar & Language Use
☐ Vocabulary ☐ Mechanics ☐ None of these components

B. Which component of the grid was the most difficult to evaluate? Please check one only.
☐ Content ☐ Organization ☐ Grammar & Language Use
☐ Vocabulary ☐ Mechanics ☐ None of these components

3A. Did you find any of the band descriptors vague or incomplete, thereby causing you to waver in your judgement of a grade? ☐ YES ☐ NO

B. Please specify which element(s) of the band descriptors you found vague or incomplete.

_____

_____

_____

_____

_____

_____

_____

Part 2: Focus on the models.

1. Did you use the models that were provided in evaluating each composition?
   ☐ YES ☐ NO

2. Did you agree with the models given to you? In other words, was the Model A paper a good representative of an A paper? Was the Model B representative of a B paper? Etc.
   Please use "X" to check your response for each model.

   Model A paper:

1	2	3	4	5	6	7
Not at all representative						Perfectly representative

Model B paper:

1	2	3	4	5	6	7
Not at all representative						Perfectly representative

Model C paper:

1	2	3	4	5	6	7
Not at all representative						Perfectly representative

Model D paper:

1	2	3	4	5	6	7
Not at all representative						Perfectly representative

Model F paper:

1	2	3	4	5	6	7
Not at all representative						Perfectly representative

3. Would your evaluation of each paper have been different if you used the grid only?
   ☐ YES ☐ NO

4. How different would your mark for each paper have been if you had not used the model papers?

1	2	3	4	5	6	7
Not at all different						Extremely different

5. Would you recommend that Credit ESL teachers be given model compositions to use in rating their students' papers?

1	2	3	4	5	6	7
Do not recommend						Highly recommend

APPENDIX D

Pre-grading Questionnaire

Before you begin grading the 16 compositions with the 5 models and the grid, I would like you to read the five model compositions attached here. I also would like you to answer the following questions on the contents of each model. You might find it strange that I am asking you these questions. However, although I am certain that you will use the models in evaluating the compositions, I need to have evidence on paper that you have actually read the models. Again, thank you so much for your kind indulgence. Your cooperation is very much appreciated.

Please use "X" to indicate your response.

1. Model A composition

A. According to the writer of the paper, physical punishment can cause a child to lack self-confidence.
□ True □ False

B. An example of "physical damage" the writer gives is
□ black eye □ broken limb □ fat lip □ none of these

C. According to the writer, a better way to discipline a child would be to
□ yell at them □ make them do homework □ make them do house work □ take away things they like to do

2. Model B composition

A. _____ is an example of "physical damage" resulting from physical punishment given by the writer of this paper.
□ black eye □ bruised kidney □ deafness □ none of these

B. _____ is a consequence the writer claims can arise from inflicting physical punishment on a child?
□ hysteria □ anti-social behaviour □ depression □ none of these

C. According to the writer, physical punishment will have no effect on a child with a "rebel personality".
□ True □ False

3. Model C composition

A. According to the writer, parents who physically punish their children
□ aren't training their children properly □ are physically abusive □ have been physically abused □ none of these

B. How, according to the writer will physical punishment "create a gap between children and their parents"?
□ children will hate their parents □ children will lie to their parents □ children will run away from home □ none of these

C. According to the writer, physical punishment can be harmful to society. □ True □ False

4. Model D composition

A. According to the writer, physical punishment can lead to _____, thus preventing a child from having a "normal" life.
□ a physical handicap □ mental illness □ crime □ none of these

B. According to the writer, physical punishment lead to sadomasochism. □ True □ False

C. According to the writer, which of the following is true about adults who physically punish their children?

□ they may go to jail | □ they may become thieves | □ they may become physically abusive | □ none of these

5. Model F composition

A. In this paper, the writer suggests that physical punishment can break the bond between a child and a parent. □ True □ False

B. According to the writer, too much physical punishment can lead to _____.

□ broken bones | □ bruised ego | □ poor communication | □ none of these

C. The writer believes that physical punishment is _____.

□ useful to a small degree | □ never useful | □ always useful | □ often useful

APPENDIX E

ESL 209, Essay Evaluation Grid Section: _____

Student's Name: _____ ID#: _____

Remarks: Strong items may be checked; weak items may be circled

Content: Ideas & Information (15%)							Thesis statement (identifiable & appropriate) A) Introduction B) Topic development: Support (body paragraphs) - on/off topic - quality- depth - relevance (unity) - fact vs. opinion - general vs. specific C) Conclusion Originality / Interest value Information value
Excellent	A+	A	A-	15	14	13	
(Very) Good	B+	B	B-	12.5	12	11.5	
Satisfactory	C+	C	C-	11	10.5	10	
Weak	D+	D	D-	9.5	9	8.5	
Fail		F		7	5	3	
					Content	/15	
Organization & Text Structure (15%)							Structure, clarity (outline implied) Coherence (sequencing): between/within paragraphs General Cohesion: effective use & variety of transitions Relationship of ideas/smooth flow Relevant pattern of organization Topic sentences (identifiable & appropriate)
Excellent	A+	A	A-	15	14	13	
(Very) Good	B+	B	B-	12.5	12	11.5	
Satisfactory	C+	C	C-	11	10.5	10	
Weak	D+	D	D-	9.5	9	8.5	
Fail		F		7	5	3	
					Organization	/15	
Grammar & Language Use (50%)							Clause & sentence structure Sentence variety Sentence problems: - fragments - comma splices - run-ons Word order Phrase structure Verb structures Articles Pronouns Prepositions
Excellent	A+	A	A-	50	46	43	
(Very) Good	B+	B	B-	41	40	38	
Satisfactory	C+	C	C-	36	35	33	
Weak	D+	D	D-	31	30	28	
Fail		F		23	15	8	
					Grammar	/50	
Vocabulary (Terminology) (15%)							Word forms Word choice (precision, suitability) Sophistication, register Idiomatic usage Variety (use of synonyms) Range (extent of word bank)
Excellent	A+	A	A-	15	14	13	
(Very) Good	B+	B	B-	12.5	12	11.5	
Satisfactory	C+	C	C-	11	10.5	10	
Weak	D+	D	D-	9.5	9	8.5	
Fail		F		7	5	3	
					Vocabulary	/15	
Mechanics (5%)							Punctuation Spelling Capitalization Paragraph form (indentations) Handwriting General appearance
Excellent	A+	A	A-	5	4.6	4.3	
(Very) Good	B+	B	B-	4.2	4	3.8	
Satisfactory	C+	C	C-	3.7	3.5	3.3	
Weak	D+	D	D-	3.2	3	2.8	
Fail		F		2	1	0	
					Mechanics	/ 5	

Reader: _____ Total: _____/100

Teacher's initials (if different from reader): _____ Score: _____/

A+	95 - 100	B+	82 - 84	C+	72 - 74	D+	62 - 64		
A	88 - 94	B	78 - 81	C	68 - 71	D	58 - 61	F	0 - 54
A-	85 - 87	B-	75 - 77	C-	65 - 67	D-	55 - 57		

ESL/CELDT Office
Updated Fall 2003, printed 7/4/2005

APPENDIX F

EVALUATION CRITERIA FOR ESL 209
GUIDE FOR COMPLETING THE EVALUATION GRID

Content: Ideas & Information (15%)

1. **Thesis statement (explicit, identifiable; appropriate to essay type or topic; predictive)**
2. **Topic development (depth and quality/originality of information)**
3. **Support (relevant, sufficient, detailed; general vs. specific support, fact vs. opinion)**
4. **Information level/value**

Excellent (A+, A, A-) Very clear and appropriate thesis, defined and supported with sound generalizations and substantial, specific, and relevant details; distinctive, original content for maximum impact; excellent information level; strong introduction and conclusion.

(Very) Good (B+, B, B-) Clear and appropriate thesis; selects; suitable and appropriate content with sufficient details; informative; occasional minor problems with focus, depth, and/or unity; good introduction and conclusion.

Satisfactory (C+, C, C-) Thesis may be unclear (e.g. too broad/narrow); acceptable topic development; some support points may be vague, insufficient, obvious, unconvincing; satisfactory introduction and conclusion.

Weak (D+, D, D-) Thesis not apparent or weak; poor topic development; lacking in substance; many support points are insufficient, irrelevant and/or repetitive; low information level; weak conclusion.

Fail (F) lacks main idea; unacceptable topic development; too vague, insufficient, unconvincing, or off-topic; not enough to evaluate.

Organization & Text Structure (15%)

1. **Presence and logical sequencing of introduction, body paragraphs, and conclusion**
2. **Use of relevant patterns of organization (related to topic or essay type)**
3. **Coherent and unified relationship of ideas (NB: grammatical accuracy related to cohesive devices is considered under Grammar & Language Use)**

Excellent (A+, A, A-) - exceptionally clear plan connected to thesis; well organized, effective and logical sequencing; smooth flow of ideas; excellent use of transition techniques; clarity of message enhanced by organization.

(Very) Good (B+, B, B-) - appropriate pattern of organization relevant to topic or essay type; generally smooth flow of ideas and appropriate use of transition techniques; overall organization good; most transitions used appropriately but would benefit from more frequent and varied use of transitions; sequencing generally logical.

Satisfactory C+, C, C-) - shows understanding of pattern of development; somewhat choppy; relationships between ideas not *always* clear; overall organization satisfactory, but some elements may be loosely connected or lacking in transitions; most points logically sequenced but some problems in organization still exist.

Weak (D+, D, D-) - problems with pattern of organization; disjointed; ideas do not flow well and relationships between ideas are often not clear; ideas difficult to follow because they are *often* not logically sequenced and/or are unrelated

Fail (F) - does not show understanding of pattern of organization; no clear organization: confusing, vague, or seemingly unrelated ideas; pattern of organization not pertinent to topic/essay type; ideas not developed in separate paragraphs; not enough text to evaluate

Grammar & Language Use (50%)

1. **Sentence structure (coordination and subordination; variety)**
2. **Sentence problems (fragments, comma splices, run-ons)**
3. **Verb structures (agreement, tense, form)**

ESL/CELDT Office
Updated Fall 2003, printed 7/11/2005
MQ

4. **Phrase structure**
5. **Articles, pronouns, prepositions**

Excellent (A+, A, A-) – sentences skillfully constructed, effectively varied with simple and complex forms; harmonious agreement of content and sentence design; hardly any errors in basic sentence or grammatical forms

(Very) Good (B+, B, B-) – sentences accurately and coherently constructed with some variety; good use of complex constructions; only a few errors in grammatical forms; meaning not affected by errors.

Satisfactory (C+, C, C-) - effective but simpler constructions and/or problems with complex constructions; meaning generally clear; several errors in grammatical forms.

Weak (D, D+, D-) - some problems in simple constructions and/or frequent problems in complex constructions, or avoidance of complex structures; clarity weakened by awkward grammatical structures; many problems in grammatical forms.

Fail (F) - many problems in sentence structures (both simple and complex) and/or absence of complex structures; frequent sentence structure errors which confuse and distract the reader; frequent errors in grammatical forms; not enough text to evaluate.

Vocabulary (Terminology) (15%)

1. **Word forms**
2. **Word choice (precision)**
3. **Register**
4. **Idiomatic usage**
5. **Range**

Excellent (A+, A, A-) high level of sophistication; impressive range; effective use of vocabulary to express ideas; only a few minor errors with word choice/form/idiom.

(Very) Good (B+, B, B-) – (very) good range and variety in the use of vocabulary; effective word/idiom choice and usage; appropriate register; several minor errors related to word choice/form/idiom.

Satisfactory (C+, C, C-) – adequate range in the use of vocabulary; occasional errors of word choice/form/idiom or usage, meaning generally clear (some minor ambiguity).

Weak pass (D+, D, D-) - limited range; frequent errors of word choice/form/idiom and usage; meaning sometimes unclear or ambiguous as a result of errors.

Fail (F) - very limited range; words recycled, reused, or too general; frequent errors of word choice/form/idiom and usage may obscure the meaning; problems with basic vocabulary; not enough text to evaluate.

Mechanics (5%)

1. **Punctuation**
2. **Spelling**
3. **Capitalization**
4. **Presentation (NB: punctuation involving fragments, comma splices and run-ons are considered under Grammar & Language Use)**

Excellent, (A+, A, A-) – very few errors either in punctuation, spelling, or capitalization; correct indentation; neat presentation.

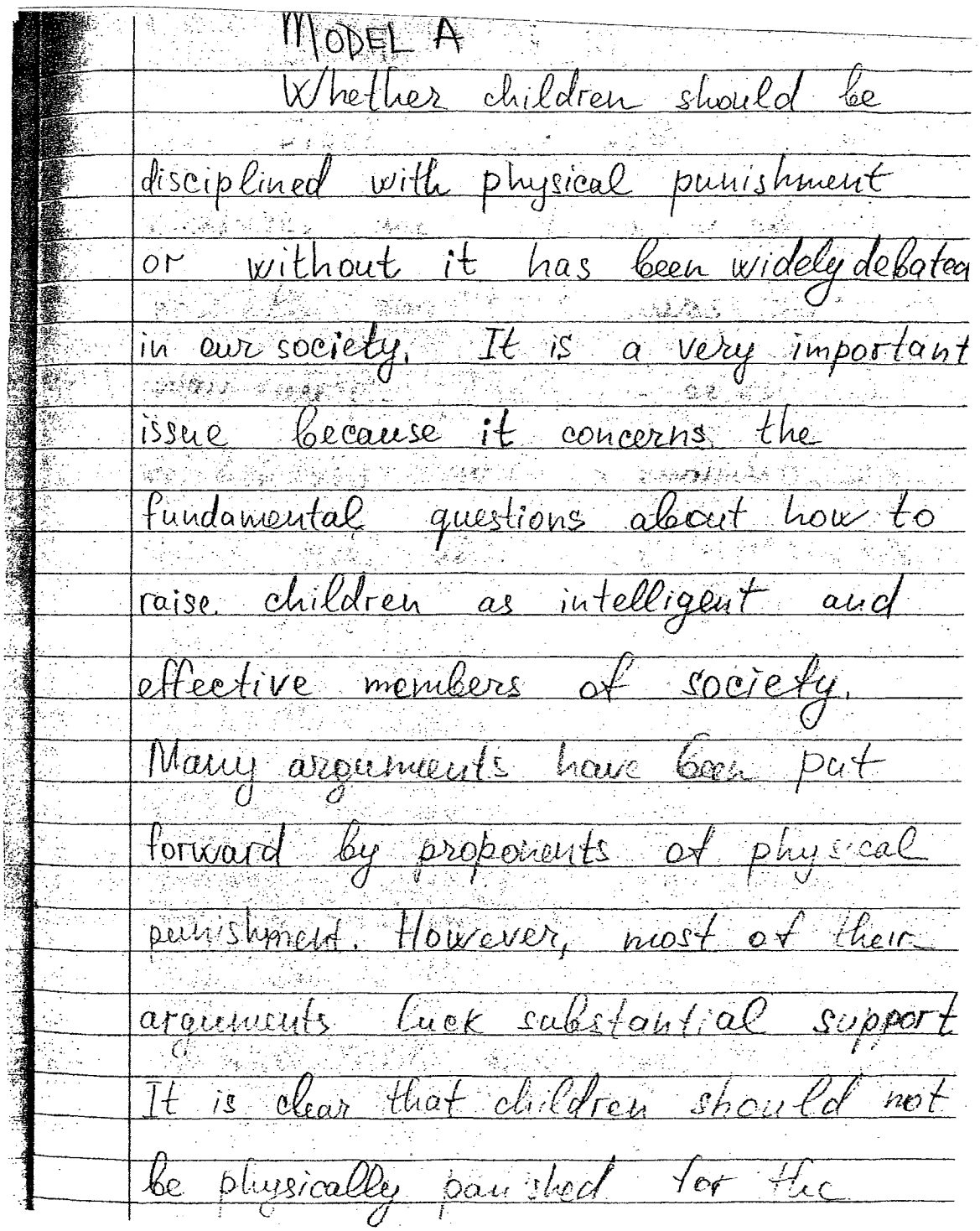
(Very) Good (B+, B, B-) - only a few minor errors in punctuation, spelling, and capitalization; clarity of message never affected by errors; correct indentation; legible handwriting.

Satisfactory (C+, C, C-) - occasional errors in punctuation, spelling or capitalization, problems with indentation; meaning still clear despite errors; handwriting hard to read but basically legible.

Weak (D+, D, D-) - many errors in punctuation, spelling, capitalization; meaning sometimes unclear as result of mechanical errors; absence of indentation; nearly illegible handwriting affecting text comprehension

Fail (F) - dominated by errors in punctuation, spelling, indentation and capitalization; illegible handwriting.

APPENDIX G



MODEL A

Whether children should be disciplined with physical punishment or without it has been widely debated in our society. It is a very important issue because it concerns the fundamental questions about how to raise children as intelligent and effective members of society. Many arguments have been put forward by proponents of physical punishment. However, most of their arguments lack substantial support. It is clear that children should not be physically punished for the

reasons that follow.

First reason that physical measures are bad for a child is possible emotional scars that it may cause. When he or she misbehaves and a parent spanks him, that child may not necessary connect the pain with his actions. That pain creates resentment in a child toward his parents because he does not understand why people that say they love him, can inflict the pain on him. In addition, that child may become withdrawn, or feel inferior. Moreover, in the future it

will result in low self-esteem.

Second reason not to spank a child for disciplining purposes is that it may cause physical damage to child's body. Parents' ~~anger with children~~ may create very high emotions, so an adult can underestimate his strength while penalizing a child. Moreover, an adult can easily break a limb or other very serious permanent harm to a small person. Therefore, obviously it is a dangerous tool for disciplining.

Some may argue that physical punishment for children is a fast and effective way to teach him or her the order, and the difference between right and wrong. However, that idea lacks substantial support. Many observation and studies conducted by child behaviourists and psychologists show that patience and love produce better results. In fact, these little people have their own "currency". These are the things that a child likes to do, for example, (it can be) watching television, or playing outside.

If every parent takes time to find out what his child likes, it will become a very effective leverage. Moreover, give children reasons to do something or to avoid some particular behaviours, and it will make him think and develop judgment.

In conclusion, physical punishment should not be used because it may cause physicall trauma in addition to psychological "scars", and, more importantly, it does not work. Love and patience are the most powerful tools.

APPENDIX H

Instructions for Grading with Trait Scale

Once again, thank you so much for agreeing to participate in my study. Please check that you have the following in your package.

- **SIXTEEN** ESL 209 argumentation essays entitled, *TC 1*, *TC 2*, etc.
- **SIXTEEN** ESL 209 essay evaluation grids, or *Trait Scales* (stapled to the essays)
- **ONE** set of band descriptors entitled, *Appendix A: Evaluation Criteria for ESL 209* (stapled to *TC 1*)
- **ONE** debriefing questionnaire entitled, *Debriefing Questionnaire—Group A*

When grading the compositions, please print your first and last name on each grid at the bottom left part of the page. Again, this is just for identification purposes. It will not be used for any other purpose. Also, please make sure that the grid corresponds to the composition.

You have **seven** days to grade the compositions and return the material, which can be left in my box at the TESL Centre. (Note: You will find a box on the desk counter in the reception room with my name on it.)

HOW TO GRADE WITH THE GRID

For each component on the grid,

a. circle the letter grade (e.g., B+, B, etc.)
b. circle the corresponding numerical value
c. check strong items featured in the grid descriptors
d. circle weak items featured in the grid descriptors
e. calculate the total score for each composition

If you have any questions, please do not hesitate to contact me. Again, thank you so much for your participation. I really appreciate your help in completing my M.A. thesis.

Marcello Quintieri
(450) 674-7658
m_quinti@education.concordia.ca

APPENDIX I

Instructions for Grading with Models and Trait Scale

Once again, thank you so much for agreeing to participate in my study. Please check that you have the following in your package.

- **SIXTEEN** ESL 209 argumentation essays entitled, *TC 1*, *TC 2*, etc.
- **SIXTEEN** ESL 209 essay evaluation grids, or *Trait Scales* (stapled to the essays)
- **ONE** set of band descriptors entitled, *Appendix A: Evaluation Criteria for ESL 209* (stapled to *TC 1*)
- **FIVE** Model Compositions entitled, *Model A*, *Model B*, *Model C*, *Model D* and *Model F*
- **ONE** questionnaire entitled, *Pre-grading Questionnaire* (stapled to *Model A*)
- **SIXTEEN** questionnaires entitled, *Grading with Models Questionnaire* (stapled to the essays)
- **ONE** debriefing questionnaire entitled, *Debriefing Questionnaire—Group B* (stapled to *TC 16*)

When grading the compositions, please print your first and last name on each grid at the bottom left part of the page. Again, this is just for identification purposes. It will not be used for any other purpose. Also, please make sure that the *Grading with Models Questionnaire* and the grid correspond to the composition.

You have **seven** days to grade the compositions and return the material, which can be left in my box at the TESL Centre. (Note: You will find a box on the desk counter in the reception room with my name on it.)

MODEL COMPOSITIONS, *PRE-GRADING QUESTIONNAIRE* AND *GRADING WITH MODELS QUESTIONNAIRE*

The Model Compositions are five essays that represent an excellent paper (Model A), a very good paper (Model B), an average paper (Model C), a below average paper (Model D), and a failing paper (Model F).

Before you begin grading the sixteen compositions with the five models, I would like you to read the five Model Compositions and answer the questions on the *Pre-grading Questionnaire*. You might find it strange that I am asking you these questions. However, although I am certain that you will use the Models in evaluating the compositions, I need to have evidence on paper that you actually read the models.

The *Grading with Models Questionnaire* contains questions about the extent to which the Models are similar to the Target Compositions. This questionnaire is to be used after you have read each Target Composition and before you grade each composition with the grid.

HOW TO GRADE WITH THE MODELS AND GRID

1. Read a Target Composition (e.g., *TC 1*)

2. As you read, mentally choose a Model Composition. Feel free to change the model as you read the Target Composition, but settle on one. Keep this model in mind as you continue reading the Target Composition.

3. After you have finished reading the Target Composition, complete the *Grading with Models Questionnaire* to decide how the paper you are grading compares with that of the model in terms of an over all mark as well as a mark for each of the five components.

4. Now, use the grid to give the Target Composition you have finished reading an over all mark as well as a mark for each of the five components. (Please circle the letter grade and corresponding numerical value.)

5. When using the grid, check strong items and circle weak ones featured in the grid descriptors.

6. Calculate the total score for each composition.

If you have any questions, please do not hesitate to contact me. Again, thank you so much for your participation. I really appreciate your help in completing my M.A. thesis.

Marcello Quintieri
(450) 674-7658
m_quinti@education.concordia.ca

PC 1

When children make something wrong, parents always want to teach children a lesson. Sometimes, parents complain that it is not useful to discipline children with words, so they have to physically punish their children. However, although I admit that disciplining is necessary, I do not agree physical punish children because it will make children lie, destroy their confidence, and cause physical trauma.

Children who are disciplined with physical punishment will tend to lie. In order to escape from physical punishment, many children will lie to their parents. Children may think that if they commit they have done wrong things, their

parents will physically punish them, then, they will choose to lie. On the other hand, if children know when they do something wrong, their parents will help them resolve the problem instead of physical punishment, they will tell the truth and ask for the help. Therefore, without physical punishment, children will really learn a lesson from their mistakes.

While children who are disciplined with physical punishment will tend to lie, their confidence will also be destroied. Since they are disciplined with physical punishment, their friends, classmates will sneer them. In their point of view, children are disciplined with physical

punishment because they are bad children.

Besides, if children are often disciplined with physical punishment will think they are bad children, always do wrong things, gratually they will lose their confidence to try new things.

The supporters of physical punishment argue that they just want to teach children a lesson, not hurt them. Actually, physical punishment will cause physical trauma. Survey shows that physical punishment will cause children's bodily or mental injury, especially children are under 5 years old. Although, parents don't want to hurt their children, they may do a wrong thing when they lose their control to

physically punish their children.

In conclusion, physical punishment will make children lie, destroy their confidence, and most importantly will cause phsical trauma. Therefore, parents should not discipline children with physical punishment and find other ways to teach children.

TC-1

Children should not be disciplined with physical punishment

Today people pay more attention on child-education. Parents want to know whether the physical punishment is a good way to discipline their children. In western countries, the physical punishment is not allowed, however, a lot of parents discipline their children with physical punishment in eastern-culture countries such as China. To agree or disagree with it is a matter of balancing between its pros and cons. In my point of view, children should not be disciplined with physical punishment because it is bad for child to grow and it will increase the number of crimes of young people.

First, we can observe easily that physical punishments are bad for a child's growth. Parents always use physical

punishments when the child does something wrong. Although the physical punishment is a good way to prevent the child to make the same mistake in the future, it kills the nature of children. A child grows by making mistakes! If they are often punished physically, they will be afraid to do anything and try anything in their lives. According to a new survey, in 2012, children with more creation ability are more living in a favorable environment which means no physical punishment. Parents should understand that the physical punishment is bad for a child's growth.

Second, the reason why I advocate that children should not be disciplined with physical punishment is that the number of physical punishments increase young people's crimes. According to a research from national crime center, about 80%

criminals between 18 to 25 were often physically punished when they were a child. When children are always disciplined with physical punishment, they generally take physical punishment as the best way to deal with problem in their own lives. That's why some young people fight each other because of some small matters. On the other hand, using good education methods let children know how to deal with problem in a proper way. Parents should understand that if they discipline their children with physical punishment, their children might treat others violently as well.

However, some people might claim there there are some advantages to use physical punishment such as the idea that it is an efficient way to warn children from dangers. Although at first glance this argument sounds reasonable

and appealing, it is not borne out by a careful consideration. There are always some better methods to teach children what to do and what not to do. Talking with children, repeating the same material, showing the children are some of methods. Parents should provide using physical punishment.

Taking all these factors into consideration, it comes conclusion easily that children should not be disciplined with physical punishment.

APPENDIX L

Grading with Models Questionnaire

Please answer the following questions after you have finished reading each composition. Please use "X" to indicate your response.

1. When you read this paper for the first time what model did you have in mind?

☐ Model A ☐ Model B ☐ Model C ☐ Model D ☐ Model F

2. What model did you finally decide to use in rating the composition?

☐ Model A ☐ Model B ☐ Model C ☐ Model D ☐ Model F

3. For each component, indicate to what extent you feel this paper is similar to the model you were using as a grading guide.

CONTENT

1	2	3	4	5	6	7
Not at all similar						Very similar

ORGANIZATION

1	2	3	4	5	6	7
Not at all similar						Very similar

GRAMMAR & LANGUAGE USE

1	2	3	4	5	6	7
Not at all similar						Very similar

VOCABULARY

1	2	3	4	5	6	7
Not at all similar						Very similar

MECHANICS

1	2	3	4	5	6	7
Not at all similar						Very similar

4. For each component, indicate whether this paper is *equivalent to*, *better than* or *worse than* the model you were using as a grading guide.

A. CONTENT	☐ equivalent to	☐ better than	☐ worse than
B. ORGANIZATION	☐ equivalent to	☐ better than	☐ worse than
C. GRAMMAR & LANGUAGE USE	☐ equivalent to	☐ better than	☐ worse than
D. VOCABULARY	☐ equivalent to	☐ better than	☐ worse than
E. MECHANICS	☐ equivalent to	☐ better than	☐ worse than

5. How confident are you about the grade you gave this paper?

1	2	3	4	5	6	7
Not at all confident						Very confident