EVALUATING AN ADDITIONAL MEASURE FOR THE AMERICAN UNIVERSITY OF SHARJAH ENGLISH PLACEMENT TEST: USING ANALYTICAL SCORING

A THESIS IN TEACHING ENGLISH TO SPEAKERS OF OTHER LANGUAGES

Presented to the Faculty of the American University of Sharjah

College of Arts and Sciences

in partial fulfillment of

MASTER OF ARTS

by

MARY ANNE JOHN

M. A. 1992

Sharjah, UAE

June 2009

We approve the thesis of Mary Anne John

Date of signature

Betty Lanteigne

26/3/09

Dr. Betty Lanteigne
Assistant Professor
Thesis Advisor

26/3/2009

Dr. Rodney Tyson
Associate Professor
Graduate Committee

P. Crompton

26/3/09

Dr. Peter Crompton
Assistant Professor
Graduate Committee

I S Sadek

26/3/09

Dr. Ibrahim Sadek
Program Director, MA TESOL

26/04/09

Dr. William Heidcamp
Dean of the College of Arts and Sciences

KMitchell

20/4/09

Dr. Kevin Mitchell
Director, Graduate and Undergraduate Programs

EVALUATING AN ADDITIONAL MEASURE FOR THE AMERICAN UNIVERSITY OF SHARJAH ENGLISH PLACEMENT TEST: USING ANALYTICAL SCORING

Mary Anne John, Candidate for the Master of Arts Degree

American University of Sharjah, 2009

ABSTRACT

The English Placement Test (EPT) at the American University of Sharjah (AUS) has been accurately placing the majority of newly admitted students into appropriate course levels, but there is a small number of students who seem to be misplaced by the English Placement Test. The issue of misplacements is a common problem with placement tests where one measure does not accurately determine placement for all test takers. It is because of this issue of misplacements that we need additional evidence to determine accurate placement levels. In this research, the first step was to investigate the use of Jacobs, Zingraf, Wormuth, Hartfield, and Hughey's (1981) analytical scale with the AUS EPT. 65 EPTs of students representatively selected in Fall 2007 were analytically double scored by two raters, and Cronbach's alpha was calculated to determine inter-rater reliability. Then, the students' EPT writing samples were evaluated by Criterion ^{SM}, a computer-based scoring program developed by the Educational Testing Service (ETS). Criterion ^{SM} and the holistic EPT were used as established measures to ascertain criterion-related validity of the analytically scored EPT in the context of this research. After establishing normal distribution and linear relationship, the analytical EPT scores were then correlated with the Criterion^{SM} scores and the holistic EPT scores, using the Pearson correlation coefficient in order to ascertain the appropriateness of use of the analytical scale in this context. These analytical EPT scores were found to be in strong correlation with the holistic EPT scores and the Criterion^{SM} scores.

Therefore in the next step of the research it was determined which of the students who had taken the EPT in Spring 2008 had been identified by their teachers as being “borderline/potentially misplaced” through classroom observation and their in-class writing samples. Once these students were identified, their EPT scripts were double scored by two raters, using the analytical scale developed by Jacobs et al. (1981), and were also evaluated by Criterion^{SM}. The placements of the identified students were re-evaluated using the analytical scale and Criterion^{SM}. Results indicated that analytical scoring with the Jacobs et al. (1981) scale and Criterion^{SM} provide valuable information complementing the holistic EPT, with a more complete profile for re-evaluating placement. The analytical scale was found to be a useful additional measure in cases of potential misplacement identified by teacher observation.

CONTENTS

Chapter Page

LIST OF APPENDICES

Appendices Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ACKNOWLEDGEMENTS

I am grateful to my advisor, Dr Betty Lanteigne, who finally made this research come true. Dr Lanteigne’s expertise in the field of language assessment is one I could not have done without. At every step, her guidance and support helped make this endeavor successful. I am grateful to my committee members, Dr Rodney Tyson and Dr Peter Crompton. Dr Tyson’s detailed comments made me rethink many a point and Dr Crompton’s suggestions helped enrich this study. Much appreciation is due to the Office of Research for providing the Seed Grant which was used to obtain the annual subscription to Criterion ^{SM} . I am also grateful to Dr Marina Dodigovic who helped give a direction to my study and who was responsible for leaving me in the very responsible and committed hands of Dr Lanteigne. I also wish to thank Dr Sarah Shono, who gave me very valuable comments during the proposal stage of this research. My sincere thanks also go out to Ms Moushumi Nandy, who very kindly conceded to be the second rater in this project. Her time and dedication to the project was truly noteworthy. I would also like to thank my chair, Dr Donald Cruickshank, and my colleagues at the Department of Writing Studies without whose support this study would have been incomplete. I cannot forget the students who allowed me to use their papers and Hadeel in the administration who helped enormously in carrying out this study. Last but never the least, I am indebted to Dr Fatima Badry for all she meant to me, not only as an efficient program director and a sound professor but more so, as a source of great inspiration and support.

DEDICATION

I dedicate this work to my mother and my late father whose prayers and blessings granted me the strength of will, courage, and determination to complete this task. I also dedicate this work to my husband and children for their patience, and for their understanding of the constraints of this study on our lifestyle for the past four years. My gratefulness extends to my brothers, who have always believed that their sister could do it even through the hardest times. May God bless them all.

INTRODUCTION

It was the last week of the semester, Spring 2008, at AUS where I was teaching WRI 001, a remedial level writing course. I was in my office with students streaming in to verify their final grades. There were about fifteen minutes left for students to come in. At this moment a student who had a very different query came in, and I asked him to sit down and tell me what it was. This is how the conversation went:

Student: Do you really think I belong to WRI 001? Is my English so weak? I have gone through this semester with zero percent effort, but I feel cheated and let down by the system in the university.

Me: You mean you did not gain anything by attending this course?

Student: Ma'am, I was so mentally disturbed that I never gave it my best. I attended class per force, and I could not imagine myself studying at such a poor level. In my previous school, my essays were read out in class. I got an A grade at the A levels. You know that any class discussion meant that I would always have a comment, but I just could not face the disgrace. I decided to disrupt the flow in class and you had to ask me to leave a couple of times but I waited till now to ask you this. Do you believe that I belong to WRI 001?

Me: Maybe, you wrote your placement test carelessly and that determined which level you were put into.

Student: Is it fair to judge a student only by a placement test?

Me: Well, I gave you an in-class assignment, too . . .

Student: But I did not write it. You made me write it the following week when add and drop week was over.

Me: I still felt you needed help with paragraphs and the organization of your ideas.

Student: Couldn't I pick that up in WRI 101?

Me: Well, I think you are a good speaker of the language and you have good vocabulary, but your writing still lacks coherence and unity. So, on the whole I feel you benefitted by attending this course.

At this comment the student left my office completely defeated, but that incident got me thinking. We had used only one measure to determine the placement of students when we needed to develop multiple measures to ensure accurate placement.

The Statement of the Problem

I recalled how challenging it had been teaching WRI 001 students in the Department of Writing Studies (DWS), as students enter their first year at the American University of Sharjah at various English proficiency levels. Some students come from the IEP (Intensive English Program) with two semesters of language exposure; some students come from Arabic-medium schools, while still others come from different boards of education which introduce English at various levels in the schools. Our university's only admission requirement is a TOEFL score of 530 or above, which will enable students to join the writing courses in the Department of Writing Studies (WRI 001 – Fundamentals of Academic Discourse, WRI 101 – Academic Writing, and WRI 102 – Reading and Writing Across the Curriculum), determined by the English Placement Test (EPT), a one-hour written test consisting of an essay based on a reading prompt, to assign students to WRI 001, WRI 101 and WRI 102. (See Appendix A, Sample EPT.) I, with other DWS teachers, have noticed that there are a few students in every class each semester who, according to their teachers, are "borderline" and/or, in the students' view, are "misplaced" into their assigned classes. A thirty-minute in-class written task based on a few guiding points – which differ for each class, depending on the teacher concerned (see Appendix B, In-class Writing) – is given by all WRI teachers at the beginning of the semester as a check against such potential misplacement. (These in-class writing samples are developed separately by all teachers with no common guidelines for the department.) However, sometimes this in-class writing assignment does not accurately document the actual writing level of the "borderline/potentially misplaced" students. As a result of this potential misplacement, teachers of WRI 001 have certain students whose language level (as observed in classroom interactions and assignments) indicates that they would be more effectively placed in WRI 101. At this point, there is no formal instrument in place to verify potential misplacement. Therefore, every semester there is a reiteration of this same problem with "borderline/potentially misplaced" students.

For the last two years, in every section of WRI 001 I have had students who thought that they would benefit from placement in WRI 101. Unfortunately, their scores on the EPT and the in-class writing assignment that I gave them during their first week in WRI 001 did not indicate a high enough proficiency level for placement in WRI 101.

These students subsequently spent the entire semester in WRI 001, even when their performance in class was at a much higher level than that of their classmates. This semester also, I have students who think their in-class writing skills indicate that they should be placed at a higher level.

This misplacement becomes a problem for teachers as well, when they have students who grasp the subject matter faster than the others, and who need to be constantly provided with more challenging work which is inappropriate for the others in the class who are generally much weaker in their writing skills and therefore slower in their written output. As a result, both teachers and students suffer; both feel stretched between two very different categories of students. Teaching to these few, more advanced students would surely neglect the needs of the majority, but not doing so would leave these "borderline/potentially misplaced" students feeling unchallenged and unproductive.

Misplacement is a common feature in language placement testing (Alderson, Clapham, & Wall, 1995) and not a problem unique to the American University of Sharjah English Placement Test (AUS EPT). This problem is because a test measures student performance on a particular occasion and may be affected by student-related concerns and/or measurement error (Brown, 2004). Many programs have a system in place by which the institution is able to identify misplaced students within the first week of class, using additional measurement instruments as evidence, as well as classroom observation (Coombe & Hubley, 2003). Other placement issues like the selection of the reading prompt, the administration of the test, and inter-rater reliability are issues that concern any institution which conducts student placement testing for the required courses (Davies, Brown, Elder, Hill, Lumley, & McNamara, 1999).

Observing such placement problems in the Writing (WRI) courses at AUS prompted me to search for a measure which could serve as an additional check for students who claim to be misplaced. Since the EPT is a holistic measure (see Appendix C, Scoring the English Placement Test), I wanted to use an analytical measure for the "borderline/potentially misplaced" students to provide additional evidence for their cases of potential misplacement. A widely used analytical scale for evaluating English as a Second Language (ESL) essays at the college level in North America (Hughes, 2003) is the composition profile by Jacobs, Zingraf, Wormuth, Hartfield, and Hughey (1981). (See

Appendix D, The Analytical Scale.) It is this scale which I am investigating as an additional measure to score the EPT essays for the "borderline/potentially misplaced" students.

To use the Jacobs et al. (1981) analytical scale to evaluate the "borderline/potentially misplaced" students, it is first necessary to validate the appropriateness of its use at AUS by correlating it with established measures which, in this research, are the holistic EPT and Criterion^{SM},˙ an online writing skills assessment program to determine writing proficiency in English developed by the Educational Testing Service (ETS). Both Criterion and the holistic EPT have similar constructs, which is why I used them in evaluating the validity of the use of the analytical scale with the EPT and identifying "borderline/potentially misplaced" students for Fall 2007. I used the analytical scale and Criterion^{SM} to re-evaluate the identified students' placement.

My original research plan was to use the holistic EPT, the analytical EPT and the in-class writing samples to re-evaluate the placements of the identified "borderline/potentially misplaced" students in Spring 2008. However, since the in-class writing samples were not available, Criterion^{SM} was used (because it was readily available although involving a subscription fee) to provide additional evidence for the issue of placements for the identified potentially misplaced students in Spring 2008. Doing so addresses three research questions in three steps.

Research Questions

Research Question 1: What does *a priori* and *a posteriori* evidence reveal about the validity and the reliability of the proposed analytical scale for use with the EPT in the Fall 2007 semester?

- *A priori* evidence for the use of the analytical scale by Jacobs et al. (1981) in this research involved considering a scale which is widely used and accepted in North American universities, and evaluates the same constructs as the holistic EPT.

- *A posteriori* evidence consists of correlating the analytical EPT scores with the holistic EPT and Criterion^{SM} scores, using scatterplot, Pearson correlation coefficient, overlapping variance and Cronbach's alpha to investigate the appropriateness of the analytical scale for use with the AUS EPT.

Research Question 2: Who, if any, are the "borderline/potentially misplaced" students in WRI courses for the Spring 2008 semester?

- Researching this question involves having WRI teachers identify the "borderline/potentially misplaced" students through their in-class writing assignments.

Research Question 3: In addition to the holistic EPT, what do the analytical EPT results and Criterion^{SM} reveal about the appropriacy of placement of the identified "borderline/potentially misplaced" students?

- Investigating this question includes reevaluating the placement of the "borderline/potentially misplaced" students.

Significance of This Research

By correlating the Criterion^{SM} and the holistic EPT scores with the analytical EPT scores, I wished to examine criterion-related evidence of validity for the analytically scored EPT in this context. This correlation was indeed important. If the scores of Criterion^{SM} and the holistic EPT scores correlated highly with the analytical EPT scores, then it would indicate that the analytical EPT may indeed be an appropriate measure of the students' writing ability.

Use of the Pearson correlation coefficient requires a well distributed sampling of two sets of interval scores in a linear relationship (Brown, 2001). Use of a scatterplot would indicate a visual representation of the relationship between the two variables (linear or not). Obtaining 65 EPT samples, selected from students in WRI 001, WRI 101 and WRI 102, provided a representative sampling. When I compared the overlapping variance between Criterion^{SM} and analytical EPT scores, or the holistic EPT with the analytical EPT scores, I observed the degree of similarity these measures had in assessing student writing, thus reinforcing evidence for the validity of the analytical EPT. Cronbach's alpha was also used to measure inter-rater reliability of the analytical scoring by two raters of the representively selected 65 EPTs. The *a priori* evidence had to be sound before undertaking the *a posteriori* evidence in Spring 2008. If the scores had shown no correlation, then this research would have ended with Research Question 1.

Based on current practice, the existing holistic EPT has been effectively placing the majority of students into the three classes of WRI 001, WRI 101 and WRI 102, but with "borderline/potentially misplaced" students, the analytical scale could be a valuable instrument to have available to the Department of Writing Studies (DWS), providing additional evidence for consideration of such "borderline/potentially misplaced" students. Criterion^{SM}'s e-rater technology is a thoroughly researched online writing program, and it is being used in this research as a criterion for criterion-related evidence of validity because it addresses the same construct as the holistic EPT (both evaluations being similar to the Test of Written English scale). Even after giving students the in-class writing assignments, teachers are not always able to clearly indicate the "borderline/potentially misplaced" students' level of proficiency; it is here that the analytical scale could be used as an alternative check in the case of these students. On some occasions, students complain to faculty and the head of DWS that they have been misplaced into WRI 001 or WRI 101. This analytical scale could come to the aid of DWS. By re-evaluating the EPT scripts with the analytical scale, DWS would only be making its point more valid if three measures (holistic EPT, in-class diagnostic and the analytical EPT) were to indicate that these students should have been placed at the levels they were in at that moment. On the other hand, if in the future a student's analytically scored EPT scores indicate placement should be otherwise, DWS could place these students as indicated. Thus by using the analytical scale, DWS could more effectively address the issue of misplacements.

These findings could also have a classroom application. The analytical scale can help students to identify their areas of weakness in writing and find out what is expected of them. This scale can also help them understand their mistakes in a comprehensive manner.

Role of the Researcher

In this research, it is necessary to understand that the researcher here functions in three capacities: a rater, a researcher and a teacher. During the first phase of this research, the EPT essays were analytically double scored, and I functioned as one of the two raters (Rater 1) involved in this exercise. (The other rater – Rater 2 – was an editor of a publishing house.) During the second stage of this research, I took on the role of the researcher to investigate the reliability and validity of using the analytical scale in the scoring of "borderline/potentially misplaced" students identified through their in-class writing. Finally, I was one of the WRI teachers involved in this study whose students participated in this research. Acting in all three roles had certain advantages and disadvantages at the same time, which I will discuss in the delimitations and the limitations of this study.

Delimitations of This Study

Placement testing presents problems worldwide (Chalhoub-Deville, 1999), and it is not an issue specific to AUS. The English Placement Test predominantly places students accurately into WRI 001, WRI 101 and WRI 102. However, there remains a small percentage of students whom teachers consider "borderline" and students consider "misplaced." Investigating an additional measure to evaluate this potential misplacement, this research does not seek to introduce a replacement for the EPT, but rather seeks to investigate whether or not the analytical scale can be used as an interim check and an additional piece of evidence in the case of these "borderline/potentially misplaced" students.

Key Terms Used in This Research

Borderline students: Students whom the teachers perceive as borderline, whether they deserve to be in a class higher or lower.

Potentially misplaced: Students who feel that their current placement is unfair.

Criterion^{SM}: An online writing assessment system designed by Educational Testing Service.

WRI: Writing course offered by the Department of Writing Studies at AUS.

In-class writing assessment: A thirty-minute in-class assessment of student writing conducted during the first week of the semester in WRI 001, 101 and 102.

Scripts: Answer papers of the students' EPTs which contain essays in response to a reading prompt.

AES: Automated essay scoring

TWE: Test of Written English

PEG: Project essay grade

IEA: Intelligent essay assessor

NLP: Natural language processing

REVIEW OF THE LITERATURE

To investigate the accurate placement of DWS students, we need to examine the existing research about placement testing, addressing issues of test purpose, validity, scoring and reliability. Since the EPT is a holistic measure, we also need to study the merits and demerits of holistic measurement. However, the holistic EPT also needs to be complemented with a different measure, and for this research I have used the analytical scale by Jacobs et al. (1981). Both analytical and holistic scales have an element of subjectivity, whereas Criterion^{SM}, a computer-based evaluation system developed by the Educational Testing Service, which also measures essay writing, is an objective measure used as criterion-related evidence of validity.

Placement Testing

The first issue to be considered is the purpose of a test, which in this research is placement testing. Placement testing, according to Green and Weir (2004), "is an area that has received comparatively little attention in language testing research" (p. 467). One plausible reason for this lack of attention could be that academic institutions do not want to spend their resources on tests which are working moderately successfully for them. Institutions therefore rightly claim that a "lack of resources constrains test practices in most settings to such an extent that the second role [i.e. research] cannot be adequately addressed" (Green & Weir, 2004, p. 468). The function of placement testing, in institutional opinion, is that these "placement tests often determine whether students need remedial instruction or were fit enough to attend an introductory course" (p. 19), as Crusan (2002) observes. If students are being placed into the right classes, institutions feel that placements have been successful.

Even though institutions have working placement instruments in place, these instruments have many administrative concerns. Chalhoub-Deville (1999) observes that "many institutions offering language programs at the post secondary level are faced with the problem of assigning to each student, in a very short time, the course that best suits [their] need" (p. 122). The assumption administrators and teachers have here is that once students are placed, they will fit into the level they are placed into. It is based on this assumption that there are sequential language courses according to the differing proficiency levels. However, as Green and Weir (2004) state, administrators "must be confident that students placed into a class will generally benefit from studying the same material in that they all enter the class at a similar point in relation to the syllabus" (p. 468). Even if the placement test does "reflect the features of the teaching context (such as the proficiency level of the classes, the methodology and the syllabus type)," as Davies, Brown, Elder, Hill, Lumley, and McNamara (1999, p. 145) state, other issues like the size of student population being given the placement test can cause administrative concerns. Davies et al. (1999) go on to state, "where large student intakes are the norm, efficiency in administration and marking is often a key consideration in the development of placement procedures" (p. 145).

Another issue that is a concern to placement testing is validation. As Hughes (2003) observes, "in the case of teacher-made tests: full validation is unlikely to be possible" (p. 33). Hughes also discusses how construct-irrelevant variants (mechanical features) like spelling and punctuation "can invalidate the scoring of written work" (p. 33), which is of crucial significance in a handwritten test. Other questions regarding the authenticity of placements and their accuracy raised by Hughes are as follows: 1) "how do we know that the placement test is measuring writing ability?" (p. 31) and, 2) can a one-hour placement test "give a sufficiently accurate estimate of the students' ability with respect to the functions specified in the course objectives?" (p. 30), when such a time limitation may result in construct underrepresentation or not fully measuring the construct in question. In most cases these placement tests are one-shot measures, implying that by this one attempt, students will be placed into different levels. As Crusan (2002) comments, "these timed writing tests . . . are given under artificial conditions" (p. 21) with no provision to refer to additional material or revise work in keeping with the process approach criteria. She further states, "Many students may find it difficult to write 'cold' on a topic that they might never have seen before and perhaps care nothing about or, even worse, know nothing about" (p. 22). This is very true in placement testing situations where students might not be remotely interested in the topic and the reading prompt provided is one they cannot relate to because they do not know what it envisages.

Selection of Prompts for Placement Tests

A third area of concern with regard to validity in placement testing is how writing prompts are selected and how certain rhetorical modes like argumentative, analytical or personal determine student performance on a test. Huot (1990a) pointed out that "researchers have often wondered whether the type of writing called for in a prompt could have a pronounced effect on the scores given to that group of essays" (p. 240). However, writing assessment research conducted so far has no clear answers as to whether or not the selection of the prompt has an impact on students' test performance. As Purnell (1982) states, institutions, rather than being affected by this issue regarding the suitability and selection of prompts, are more concerned with "getting clear and consistent standards for passing the test" (p. 409), and it is this lack of attention given to the suitability and selection of prompts that "offers the most serious challenge to all involved in the testing process" (p. 410). Institutions, in addition to wanting clear standards, also want these standards at the minimum price. Green and Weir's study (2004) confirmed that "low-cost placement instruments may provide crude indications of student abilities but, unless supported by more extensive procedures, they are unlikely to provide the kind of detailed diagnostic information that is desirable for teachers and learners" (p. 488).

Detecting Problems in Placements

It is important for institutions to acknowledge whether or not they have discovered problems with their placement instrument – problems like misplacing students in classes or miscalculations on the part of the assessors resulting in a wrong score. As Alderson, Clapham, and Wall (1995) observe, "the fact that the testing body has made a mistake will not make a bad impression on the teachers or students if the testers make it clear that they treated the candidates fairly in the end" (p. 205). Institutions need to recognize the importance of placement decisions on instruction. According to Haswell and Wyche-Smith (1994), the "emphasis of placement should be less on scoring reliability and more on instructional validity" where the majority of students "are quickly and reliably placed, and only a few recalcitrant pieces require thorough analysis" (p.

692). In order to develop a valid test, as Bachman (1991) states, institutions need to specify the “characteristics of the test task and the test taker’s language ability” (p. 692). Institutions should, as Bachman advised, “design and develop language tests that are potentially more suitable for specific groups of test takers and more useful for their intended purposes” (p. 677). Bachman goes on to state that placement tests should have “situational authenticity” (p. 690) where the test has a relation to the situation where the target language will be used and “interactional authenticity” (p. 691) where the test has a relation to the test taker as well. Language teachers and testers are concerned with this kind of authenticity, according to Bachman, because they want to do their best to make their teaching and testing relevant to their students’ language use needs. Thus placement testing has many issues connected with its design and use.

Scoring Placement Tests

Using Multiple Measures for Placements

Clearly, it would be a better option to rely on more than one measure for the purposes of appropriate placements. According to Coombe and Hubley (2003), we need to rely on “multiple measures of assessment” (p. 22). Such an approach seems to be a more comprehensive system of placement, since multiple measures and placements do not rely on a one-shot approach by which students are judged wholly on the basis of their performance on one given day. The use of multiple sources of information in designing and selecting assessments is also a key factor in interpreting placements effectively. This also increases “the collective reliability of the decision made,” as Brown and Hudson (1998, p. 671) state.

“Tests are neither good nor evil in and of themselves. They are tools,” according to Brown and Hudson (1998, p. 672), which need to be used considering the student body and the institution in which they are being used. Finally, as Purnell (1982) states, while placement testing “may provide certain kinds of information it should never be regarded as an infallible index to anyone’s level of competence. Nor should a test be the sole determiner of students’ future steps or of the opportunities offered to them” (p. 410). Basing placement decisions “on a single source of information is dangerous and even foolish" as Brown and Hudson (1998, p. 670) believe. Thus when considering the issue of placement tests, institutions need to understand that these tests have far greater significance for their test takers. The second language specialists in charge of placement testing need to examine placement practices periodically and "aim for a reconciliation of these practices" with their classroom pedagogies, as Crusan (2002, p. 17) observes. With regard to placement testing, these second language specialists know what they have to do but often do not do it because the "reasons are legion: cost, speed, practicality and efficiency, validity and reliability" (p. 18), as Crusan aptly states.

Holistic Scoring of Placement Tests

Placement tests are usually holistically marked to save time and for the convenience of institutional administration. These tests are usually conducted in the beginning of the semester, or earlier even, to determine which students go into what levels. At this busy time when freshmen are getting acclimatized with the college environment, and the institution also has innumerable problems, placements are yet another concern. But not only is holistic scoring easier for placement purposes, it is also more "economically feasible" (White, 1984, p. 402). It is far easier for a rater to go through student essays and come up with a general overall score rather than follow the time-consuming method of analytically scoring scripts on an elaborate set of guiding principles. Holistic scoring does, as Madsen (1983) observes, give us "one of the best ways to evaluate the complex communicative act of writing" (p. 122). As White (1984) observes, the "holistic scoring of writing samples could take place quickly enough to be practical" (p. 402). Similarly, Huot (1990a) states that grading holistically is quick and is "usually recommended especially for large testing populations" (p. 238). He goes on to comment that "holistic scoring reflects the general impression of the quality of a piece of writing" (p. 238). The common scales used to holistically mark scripts are usually the four-point scale or the six-point scale, according to Huot (1990a). The point on the scale corresponds with the quality of the test taker's writing. It is surprising to note that as far back as 1982 students preferred the use of holistic scoring as the primary method of evaluating placements, as a student survey conducted by Purnell (1982) shows. He notes that "eighty-nine percent of the respondents overwhelmingly favored the reader's general impression of the overall quality of a piece of writing over its discrete features" (p. 408). Hamp-Lyons (1990) considers placement tests as direct tests and also opines that writing samples must be considered as "whole discourse" (p. 6). White (1984) states that writing, like reading, is an exercise for the whole mind and includes its most creative and imaginative faculties. He says that

> writing must be seen as a whole, and that the evaluating of writing cannot be split into a sequence of objective activities, holisticism reinforces the vision of reading and writing as intensely human activities involving the full self. (p. 409)

However, this does not imply that holistic scores can be treated as "an absolute value" as White (1984, p. 406) observes.

Sometimes a student's essay is "internally congruent" as Hamp-Lyons (1995a, p. 760) states, and the qualities of writing may be adequately represented by a single score for large-scale testing purposes. But sometimes, as Hamp-Lyons further states, a text is so

> internally complex that it requires more than a single number to capture its strengths and weaknesses. Readers do sometimes identify and need to separate out features of essays they are trying to score in order to make sensible judgments. (p. 760)

Reid (1993) believes that though holistic scoring assesses the overall competence of a piece of writing, "it neither diagnoses problems nor prescribes remedies" (p. 235). Hamp-Lyons (1990) observes that direct tests may be useful in determining student placements into writing programs, but they have limitations with regard "to their unreliability and in part to the ways in which they were scored" (p. 2). Different raters may choose to focus on different aspects of the written product (Nakamura, 2004). However, as White (1984), referring to a survey conducted by Purnell in 1982, states, the problem with this general impression scoring was clear: "a paper's score depended on the accident of who wound up as reader rather than on its quality, however that quality was defined" (p. 403). White cites a study conducted by Diederich of ETS, renowned for his scholarship a generation ago, and refers to how different raters mark different categories differently. For example, a rater who places more emphasis on structure would score papers differently from another rater who values style as the predominant criterion. Even when holistic rating scales are composed of multi-feature level descriptors, these descriptors, as Connor-Linton (1995) observes, "risk forcing potentially multi dimensional rater responses into a single dimension of variation" (p. 763). ESL writers show varied performance on different traits, and if we do not score these traits and report these scores, much information will be lost (Hamp-Lyons, 1995a). Thus the most important limitation of the holistic score, according to White (1984), "is that it gives no meaningful diagnostic information beyond the comparative ranking it represents" (p. 406).

According to Huot (1990b), holistic raters are mostly influenced "by the content and organization of a student's writing. Raters are also most influenced by the text" (pp. 207–208). He goes on to state that "one major weakness is that little attempt has been made to analyze the process of reading and rating student writing in a holistic scoring session" (pp. 207–208). Huot believes when raters score papers they are checking whether the student is fit enough for their courses and are looking for "symptoms" in the text. He continues saying, "placement scoring may be conceived as symptomatic scoring where raters consider the courses students must take in relation to the quality of the texts they are scoring. Thus placement scores produce rankings that are site-specific and non transferable" (p. 208).

Another unique feature in placement testing is that the "raters are aware that their decisions have a direct impact" on the student's placement because rating decisions are used to place students into various courses (Huot, 1990b, p. 208). However, raters may not be sensitive to particular textual features as they concentrate on the general impact of a text. As Huot (1990b) observes, this is one of the drawbacks of holistic scoring when compared to analytical scoring which can "determine usage differences in student writing" (p. 209). In Hamp-Lyons' opinion (1995a), "a holistic system is a closed system offering no window through which teachers can look in and no access points through which researchers can enter" (pp. 760–761).

Finally, "the choice of a scoring method is not always easy" for any institution (Nakamura, 2004, p. 45). According to Huot (1990b), "any choice of holistic scoring or any other evaluation instrument should only be made in terms of the purpose and content of the specific assessment situation" (p. 209). Despite the potential problems and valid concerns of holistic scoring, it has become fairly standard practice and is now being frequently used in placement exams (Morris, n.d.). As Huot observes, “clearly the question of validity is complex and variable, depending upon the specific function of individual scoring sessions and we need to explore the various ways holistic scoring is used to establish theoretical soundness” (p. 210). Huot concludes, there is a lot we need to know about the major uses of holistic scoring and “the effects of holistic scoring practices upon the ability of raters to read and rate student writing successfully” (p. 211).

Bachman and Palmer (1996), on careful observation, found “with global scales, there is always the possibility that different raters (or the same rater on different occasions) may either consciously or unconsciously weigh the hidden components differently in arriving at the single rating” (p. 210). This point brings us to intra- and inter-rater reliability issues in holistic placement tests where scorers may not be consistent in their rating of papers, thus affecting the placements of the students concerned. Harmer (2001) states that though global assessment scales are “predefined descriptions of performance,” they fall “short of the kind of reliability we wish to achieve” (p. 329).

Analytical Scoring of Placements

Harmer (2001) believes that analytical profiles are “more reliable when a student’s performance is analyzed in much greater detail” (p. 330). Bachman and Palmer (1996) share the same opinion as they believe that analytical scales are more explicit and so “we are in a position to control the weighting of the different components, either through rater training or through statistical procedures” (p. 210).

When raters score papers analytically, Nakamura (2004) says, “they are required to focus on each of the various assigned aspects of the writing sample, so that they all evaluate the same features of a student’s performance” (p. 45). Analytical scoring “focuses on several identifiable qualities germane to good writing,” according to Huot (1990a, p. 238). He also says that these qualities are then identified and the quality of the paper “is judged by how many components of good writing it contains” (p. 238). Hamp-Lyons (1995b) states that for over a decade she has been developing what she terms “multiple trait scoring instruments” (p. 453) because she prefers that term “to move us beyond the baggage that the term analytical scoring" carries (p. 453). She goes on to state that "multiple trait scoring within a trait evaluation" (p. 454) can display the features that researchers value in holistic scoring. Hamp-Lyons (1995b) believes that "multiple trait scoring allows us to go further because it forces the construction of local theory of what good writing is, and it forces us to be specific about what we like and don't like in a text and why" (p. 454). She further states that the multiple trait assessment process is invaluable in diagnosis and complex placement issue decisions. It also "opens up to researchers all aspects of test development and operation" (Hamp-Lyons, 1990a, p. 761) together through providing detailed data.

White (1984) observed that analytical scoring "imagines a model of writing that is neatly sequential and comfortably segmented " (pp. 407–408) but there are limitations to analytical scoring. Analytical scales can also prove to be unreliable. As Madsen (1983) states, "a major problem in analytical approaches is that one never knows just how to weight each error or even each area being analyzed" (p. 121). Harmer (2001) reiterates this concern when he comments, "however well we have balanced the elements in our test, our perception of our students' success or failure will depend upon how many marks are given to each section of the test" (p. 328). As White (1984) notes, "in theory, analytical scoring should provide the diagnostic information that holistic scoring fails to provide and in the process yield a desirable increase in information from the writing sample" (p. 407). However, there are numerous problems and limitations with analytical scoring. In reality, White (1984) further notes, "analytical scoring tends to be quite complicated for readers which leads to slow scoring which in turn leads to high costs" (p. 407). He comments that analytical scoring "assumes that writing can be seen and evaluated as a sum of its parts" (p. 407) which opposes the principles on which holistic scoring is based. So, in his view though "analytical scoring offers some valuable adjunct measures with regard to some skills, it is not a valid measurement of writing" (p. 408).

Comparing Holistic and Analytical Scoring

Thus with both holistic and analytical scales there can be problems. Both have reliability and validity concerns which need to be examined at length. Reliability is a factor that is concerned with the consistency of scores, given that the test is repeated under similar conditions. As White (1984) states, “since reliability is in a sense a technical term to describe fairness, or simple consistency, good testing practice aims for the highest reliability that can be reached” (p. 403). A way to make both holistic and analytical scales reliable is to have more than one scorer. As Harmer (2001) states, “the more people that look at a script the greater the chance that its true worth be located somewhere between the various scores it is given” (p. 329). Thus with reliability is tied up the issue of who marks the papers, how many scripts they mark, and what credentials or training they have, among other factors. Also, it is safer to use more than one performance instrument to enhance the reliability of placements, as Norris, Brown, Hudson, and Yashioka (1998, p. 3) suggest when they refer to “other methods of gathering information” (namely classroom behavior) or “alternatives in assessments” that should be considered in order to make placements more reliable.

Concerns of Reliability

Issues like the clarity of the prompt, the cultural biases apparent in a prompt, or the clarity of instructions before students write the placement exam are some issues that may make a test unreliable. Even not having enough copies or not photocopying all the sides accurately can make a test unreliable. Garcia-Mayo (1996) analyzed many factors that come into play when we consider how reliable a test is. To make a test more reliable he suggests not having students write their name atop their papers to avoid any discriminatory factors in their name that raters may be able to recognize and then undermark or overmark as the case may be. He further states that students’ papers should be typed “to avoid the handwriting factor which is the most tangible source of unreliability and invalidity in essay tests” (p. 57). Topics “should be carefully phrased” (p. 58) so as to convey no ambiguity to the students who are answering them under placement test conditions. Research has proven that student performances vary “from topic to topic” (p. 57) and therefore test developers should be careful while designing placement instruments. Finally, to make testing as objective as possible, Garcia-Mayo recommends that “trained professionals” (p. 57) grade placement tests. However, many institutions find it difficult to comply with these issues because of practical and logistical considerations like the time given for marking, the assurance of secrecy while results are being compiled, and the added payment costs which professional organizations will claim for their expertise.

Also, institutions find it increasingly challenging to develop tests which are both reliable and valid. Even a professional testing agency like the Educational Testing Service (ETS) has encountered many hurdles in the pursuit of developing a sound writing test. As White (1984) notes, "the problem of developing valid, reliable and economical measures of writing ability has been a particularly difficult and thorny one for ETS" (p. 401). When the focus is on getting reliable results, the process of developing these tests suffers. Connor-Linton (1995) states, "the focus on reliability has emphasized the product of assessment, much less research has been devoted to the process" (p. 763).

How reliable can a test be when it is conducted in an unfamiliar environment? Students who have newly joined the institution, probably sitting in a new place for the first time, will be affected by this unfamiliarity. Students will approach placement tests differently than classroom tests, according to Garcia-Mayo (1996), who states that "their physical condition or even their psychological condition is something over which we have no control" (p. 53). The best we can do, he says, "is to avoid the alleged lack of reliability of the holistic method in this area is to get the writers involved in their task" (p. 53).

Not only does the placement test setting disturb the student, but the test content also disturbs the evaluators, as Huot (1990b) explains:

> Just as a writer's work is affected by their attitude to their subject matter, so too the meaning and impact of a text is controlled by the reader's purposes and any biases or feelings generated about the material they are reading. (p. 210)

This implies that the placement test content affects both the test taker and the evaluator. Also, research is inconclusive about whether or not evaluators score the way they think they do. Huot (1990a) goes on to state,

> how raters score papers seems to be a more interesting area of inquiry. The notion of whether or not raters score papers the way they think they do has not been explored fully, but the limited attention it has received has produced some contradictory results. (p. 256)

While it is agreed that holistic procedures are used by institutions for practical purposes of handling large numbers, there are many variables that go into how a particular grade is assigned. As Morris (n.d.) notes, “if the reader just finished a poorly written paper, the next one may seem exceptionally good even if it was just moderate. Likewise the first paper read could be held to a higher standard than later papers because the grader recognizes the overall performance of a class or group” (¶ 3). Garcia-Mayo (1996) reiterates this concern when he states that “holistic evaluation can be highly general and subjective due to bias, fatigue, previous knowledge of the student and shifting standards from one paper to another” (p. 53). It is therefore essential to have established criteria that carefully determine grades so that maximum reliability can be ensured.

Inter-rater and Intra-rater Reliability

Bachman (2000) states, “with the renewed interest in performance assessment” has come the “increased focus on the role of the raters in the assessment process” (p. 11). Raters are affected by two factors: inter-rater and intra-rater reliability. Inter-rater reliability refers to the differences among raters in the way they score papers, while intra-rater reliability refers to the inconsistencies in a single rater’s scoring (Bailey, 1998, p. 247). Sometimes raters can mark the same paper differently at different times of the day, depending on their mood and their attention span, while other factors can be attributed to noticing and focusing on different aspects of the writing, especially if the papers are holistically marked.

Inter-rater reliability: According to Eckes (2008), “rater variability or inconsistency between raters” (p. 156) may be due to a variety of factors. Eckes explains that raters might differ on how much “they comply to the scoring rubric” (p. 156). Raters might also have varying opinions in the way that “they interpret the criteria employed in operational scoring sessions” (p. 156). They might also differ in the “degree of severity or leniency exhibited when scoring examinee performance” (p. 156). Eckes concludes that raters might also have differences in opinion about the “use of the rating scale categories” and finally that “raters might differ in the degree to which their ratings are consistent across examinations, scoring criteria and performance tasks” (p. 156).

Intra-rater reliability: As Garcia-Mayo (1996) states, “the same composition grader may assign the same composition to different grading categories at different times, affecting intra-rater reliability” (pp. 53–54). This point of intra-rater reliability has been further discussed by White (1984) who states that “if papers were re-read it does show that most of the essay scores will change slightly upon re-reading, and that some scores will change a great deal” (p. 407), indicating that intra-rater reliability is definitely a key issue in determining accurate placements. Adding to this point is Garcia-Mayo’s (1996) statement that “research has proven that English teachers vary in their assessment of writing proficiency” (p. 55).

Inter- and intra-rater reliability: However, institutions neglect to consider this inter- and intra-rater reliability because of time constraints and pressure from the stakeholders to carry out quick placements. As Huot (1990b) states, “by necessity, holistic scoring emerged as a primary practice solely on the strength of its inter-rater reliability coefficients. When agreements between raters got high enough, usually .7 or better, holistic scoring techniques were accepted as a viable way of evaluating writing quality” (p. 204). Huot further states that this very emphasis on reliability which made holistic scoring procedures acceptable is what “has retarded their scope and stunted their growth” (p. 204) because, as he explains, “often reliability is the only consideration of test administrators, researchers, and composition scholars” (p. 204). This is why White (1984) observes that “in holistic scoring, reliabilities are customarily over estimated and the inescapable inaccuracy of scores tends to be ignored” (p. 407). When raters have to grade scripts with a certain general rubric and measurement, this alters their judgments to a certain extent, according to Huot (1990a), since “an evaluator who must judge a text by certain pre-ordained criteria and agree with his or her fellow reader cannot but alter the point of view by which a text is read” (p. 256). In other words, adhering to a rubric compromises the uniqueness of each rater, but it is this standardizing aspect that makes grading consistent among different raters using a rubric.

The best way to solve this problem of intra-rater and inter-rater reliability concerns is to provide adequate rater training. As Garcia-Mayo (1996) says, when raters are properly trained, then “they will have to agree on aspects they are going to consider while grading, a specific set of values common to all raters has to be established” (p. 56).

Huot (1990a) is of a similar opinion when he says that "scoring procedures act as a controlling influence on the disparate impact of personal experience, variation and expectation. Members of a rating session come together for a set purpose and develop a particular point of view through which they read and rate essays" (p. 257). Garcia-Mayo (1996) believes that if rater reliability is to be increased, then "proper training of those involved in the grading process should be considered and the setting of common standards and possible methods of concealing student identities should also be taken into account" (p. 58).

Rater Perception

Sweedler-Brown in a study conducted in 1985, found that the more experience and training a grader had, the lower both holistic and analytical scores were. She states that "experience and training in using the holistic criteria scale may give graders the confidence to grade more critically. Inexperienced graders are likely to be more uncertain of exactly how to interpret and apply the holistic scale" (pp. 54–55). In discussing Sweedler-Brown's (1985) study, Morris (n.d.) notes that raters who were trained and experienced tended to be more consistent in assigning grades, "but that those grades tended to be lower than the inconsistent novice grader" (¶ 4).

Though many issues remain with regard to rater perception, Huot (1990a) states:

> It is safe to say that questions about the influences of writing quality on rater perception are far from being answered, and there still remain many unanswered questions about how direct writing evaluation procedures affect rater ability to judge writing quality.
> (p. 257)

In this scenario how do we ensure optimum reliability? The answer to making placement reliable is that it must be based on multiple sources of information, according to Brown (1997): "Like qualitative researchers, quantitative researchers should stress the importance of multiple sources of information, especially in making important decisions about students' lives" (p. 17). Nakamura (2004) further suggests that "test developers develop an appropriate balance among Bachman and Palmer's (1996) six qualities of test usefulness" (p. 45).These landmark qualities developed by Bachman and Palmer are reliability, construct validity, authenticity, interactiveness, impact and practicality, by setting minimum acceptable standards.

Validity of a Placement Test

The next question to be answered with respect to placement testing is the issue of the validity of the placement instrument. Huot (1990b) states, "since the issue of inter-rater reliability and holistic scoring has been settled the profession now needs to consider the many unanswered and unasked questions about holistic scoring validity" (p. 204). As White (1984) goes on to state, "in order for the direct measurement of writing ability to become an accepted component of writing testing it has to meet the two criteria of reliability and economy, without losing its face validity as a legitimate test of writing skill" (p. 402). Validity has many facets, and we need to understand these different types of validity in order to understand how they all have a bearing on this research. The different types of validity are face validity, content validity, criterion-referenced validity, concurrent validity and predictive validity.

Face Validity

Most placement tests have face validity because they have the appearance of a test (Kroll, 1990). They are conducted in an environment that is conducive to testing. The test paper inherently has in it issues which blend with testing, as the instructions on the test paper and the marking scheme both tell the students the importance of the test. The fact that there is a fixed allotment of time and there are invigilators in the room all make the students know and feel it is a test. Huot (1990b), referring to writing tests, agrees with this when he states that "the use of student writing to assess writing ability has face validity" (p. 204).

Construct Validity

Though the placement environment is conducive to test taking, the main question here is whether or not the test measures what it is supposed to measure, or in other words, does this test have construct validity (Larsen-Freeman, 1985)? Kroll (1990) speaks of the same issue of validity when she states that "if writing tests are to do more than permit, crude, short-term decisions about who goes into which writing class, we need to ensure that tests are construct-valid" (p. 71).

According to Davies et al. (1999), "Construct validity of a language test is the indication of how representative it is of an underlying theory of language learning. Construct validation involves the investigation of the qualities that a test measures, thus providing a basis for the rationale of the test" (p. 33). In a placement situation, construct validity would imply that those students performing well are competent writers, while those performing poorly would be incompetent, according to Huot (1990b). The American Psychological Association (APA) publication of *Standards for Educational and Psychological Tests* states that construct validity can only be measured over a series of testing situations with the same testing measure (as cited in Huot, 1990b). When comparing construct validity between holistic and analytical scales, Nakamura (2004) states that though raters assume that all relevant aspects of writing ability develop at the same rate and can thus be captured by a single score, he believes it is more appropriate to rate L2 (second language) writers on different aspects of writing, as this ability "develops at different rates in different writers" (p. 46). This view implies that though raters would like to give test takers a single score for a writing task that measures different aspects of their writing, this single score will not be able to give an accurate estimate of the test takers' writing abilities as different test takers will perform differently on different aspects of the writing task. Thus analytical scoring of writing ability is more valid in terms of construct validity as it measures the various constructs on which a test is based.

Content Validity

Kroll (1990) says that a writing placement test does achieve content validity as test takers are free to select the content themselves and put forth an argument "to capitalize on what they know while underplaying those aspects of content where they are lacking" (p. 72). This, of course, depends on the specific environment of each institution and each placement test. Content validity, as defined by Davies et al. (1999), is "a conceptual or non-statistical validity based on a systematic analysis of the test content to determine whether it includes an adequate sample of the target domain to be measured" (p. 34). The target domain that the test seeks to measure could be the whole level of language or it could be a specific domain that the institution wishes to test. Thus placement tests can achieve content and construct validity but have to consider student writing samples pieces of evidence which are "not valued in themselves but only as indicators of how a person would perform similar or related tasks in the real world setting of interest" (McNamara, 2000, p. 8).

Criterion-related Evidence of Validity

One aspect of evidence for validity is criterion-related evidence of validity, which enables a new test to statistically establish a correlation in terms of the closeness of the test to an established measure. [This is different from criterion-referenced testing which, though close in terminology, is related to the use of test scores. If a test is used to rank students into various categories or levels based on their scores then it is said to have a norm-referenced system of marking. In contrast, to check whether or not students have understood certain points (or criteria), then criterion-referenced testing is done so that a rater can see how well the students have performed with reference to the elements being measured (the criteria). In placement testing, this criterion-referenced testing can help determine curricular alignment. As Cohen (1994) states, "criterion referenced assessment produces information that is more descriptive and addresses absolute decisions with respect to the institutional goal" (p. 25). He further states that this approach is used to figure out curricula alignment or whether or not a test taker "has met certain instructional objectives or criteria" (p. 26).]

In criterion-related evidence of validity we compare a new test's performance against an established test to ascertain the validity of the new test. As Davies et al. (1999) explain, "criterion-related validity of a new test is established statistically (using correlation) in terms of the closeness of a test to its criterion" (p. 39). The criterion may be an established test or another measure; "in both cases validation is judged in terms of how closely the new test correlates with the criterion measure" (Davies et al., 1999, p. 39). Tied up together with this type of evidence of validity is concurrent validity and predictive validity.

Concurrent and Predictive Validity

According to Davies et al. (1999), concurrent validity is "concerned with the relationship between what is measured by a test (usually a newly developed test) and another existing criterion measure" (p. 30). Huot (1990b) states that concurrent validity "is the ability to correlate one type of testing with another" (p. 206). Predictive validity, on the other hand, as Lloyd, Davidson and Coombe (2005) observe, "measures how well a test predicts performance on an external criterion" (p. 189). The main purpose of predictive testing is to provide information about how the test taker will perform in the real world.

Both concurrent and predictive validity are essentially two kinds of criterion-related validity. According to Hughes (2003), "concurrent validity is established when the test and the criterion are administered at about the same time" (p. 29). Hughes gives an example of a test with a ten-minute oral component. He measures this with another test which has a forty-five minute oral component. He then compares the two tests concurrently by calculating a validity coefficient, in which a perfect agreement between both sets of scores will result in the coefficient being 1 and a total lack of agreement will result in zero. Hughes continues saying that "whether or not a particular level of agreement is regarded as satisfactory will depend upon the purpose of the test and the importance of the decisions that are made on the basis of it" (p. 28).

The second aspect of criterion-related validity is predictive validity, which is concerned with the degree to which a test can predict a test taker's performance in the future, for example, when a proficiency test can predict how well or how badly a student will do in a particular course. This, of course, depends on the choice of criterion measure, and Hughes (2003) feels this raises interesting issues like, what is the criterion? Is it a supervisor's subjective judgment, or is it based on other factors like subject knowledge, intelligence and motivation, for instance?

Another example of predictive validity would be an attempt made to validate a placement test. Placement tests attempt to predict the most appropriate course level for a student. Hughes (2003) gives an apt explanation for validating a placement test. He states,

> Validation would involve an enquiry, once courses were under way, into the proportion of students who were thought to be misplaced. It would then be a matter of comparing the number of misplacements (and their effect on teaching and learning) with the cost of developing and administering a test that would place students more accurately. (p. 30)

When new tests need to be validated for placement testing, they have to be measured against prevailing established measures. Once these measures are correlated, and the success of the new test is examined, only then is the process complete. Based on results obtained, the new test may have to be modified and then put into practice as a new test, providing a sounder basis for accurate placements.

Evaluating Analytical Scales

Understandably, any research into placement testing is a complex process, and choice of an analytical scale is a complex decision. Avenues, an online program for ESL writers designed by Moser (2008), presents an analytical writing assessment rubric that could be used to assess five components of writing: focus and coherence, organization, development of ideas, voice, and written conventions. Each of these five components on the scale could be scored from 1 to 4 and could be adapted to give separate scores totaling between 5 and 20. Different researchers use different terms to imply very much the same criteria for writing evaluation. For example, O'Malley and Pierce (1996) divided the domains of writing into composing (broken down further as focus, organization and elaboration), style (similar to Avenue's concept of voice), sentence formation, usage and mechanics. Sasaki and Hirose (1996) broke down their rubric into "content, organization, vocabulary, language use and mechanics" (pp. 145–146). Earlier, Harris (1969) categorized the rubric by content, form, grammar, style and mechanics. Any one of these trait-based writing assessment rubrics could be used to measure learners' development in English writing in a way that provides students with meaningful feedback and diagnostic information (Moser, 2008).

However, in considering any rubric, it is important to remember what Clark (2003) stated: "Rubrics are generally less objective in practice" (p. 216). If a rubric calculates points from 1 to 6, Clark observed, “all the papers could be error free, but the higher scores do not indicate whether or not they allow for errors. . . . If the grader has been assigned a good sample of each of these scores and uses them as a reference when grading essays, it becomes more clear what grades should be assigned to which essays” (p. 216). Very often it happens that some scorers may be tougher than others on allocating grades. So a high grade need not necessarily imply that the student has a higher level of proficiency compared to another student who received a lower score from a stricter rater. Only when raters are given representative samples of writing for each score level can the rating process seem more reliable and uniform, because all raters will be looking for the same recognizable features in all the scripts. As Connor-Linton (1995) states, “understanding the rating process is also a prerequisite for principled improvement of rater training” (p. 764).

Computerized Scoring of Placement Tests

Both holistic scoring and analytical scoring contain a human element which makes them subjective to a certain extent because it depends on rater factors, but Dodigovic (2005) mentions that “computers have clear efficacy advantages over human rating” (p. 105), which makes them objective and thereby negates the subjectivity of human rating. Many have preferred the use of computers over human beings to score papers, as Attali and Burstein (2006) note: “Surprisingly for many, automated essay scoring (AES) has been a real and viable alternative and complement to human scoring for years” (p. 3). This is probably due to the vast advancements in technology. As Bachman (2000) states,

> On the practical side advances in the technology of test design and development along with the availability and use of ever more sophisticated computer and web-based applications for test administration, scoring and analysis, have resulted in a greater range of test formats and assessment procedures than has ever been available. (p. 2)

Automated Essay Scoring (AES)

It is important to understand how an AES system works before looking at its distinct advantages and drawbacks. AES systems do not actually read and understand essays as humans do. According to Attali and Burstein (2006), "AES use approximations or possible correlates of these intrinsic variables. This is why there has been skepticism and criticism over the years related to the fact that the machine does not understand the written text" (p. 4). The latest version of e-rater (which is used by Criterion^{SM}), according to Attali and Burstein, includes "measures of grammar usage, mechanics, style, organization, development, lexical complexity and prompt-specific vocabulary usage" (p. 7), which is feedback that students receive when they input their papers into Criterion^{SM} , ETS's writing instruction application. Also, e-rater v.2 is used by Graduate Management Admission Test (GMAT) as a second rater, which is a safe way of using any software, because it is complementing human scoring (Attali & Burstein, 2006).

Ben-Simon and Bennett (2007) say that AES systems can be classified into two categories: brute-empirical and hybrid methods. Brute-empirical implies that the computer is looking for discrete point features of writing and not at content, organization or style. Hybrid methods assess more like the way human beings do with content, organization and style being principal in determining the computerized score. The brute-empirical system extracts, as Ben-Simon and Bennett further state, "a large variety of linguistic features from an essay response which may not necessarily have any direct explicit link to writing theory" (p. 6). On the other hand, hybrid methods are "more closely related to a theoretically derived conception of the characteristics of good writing" (Ben-Simon & Bennett, 2007, p. 6).

Ben-Simon and Bennett (2007) give a brief history of AES systems. They say that Project Essay Grade, or PEG, in 1966 was the first automated essay grading system to be conceived. It made many improvements until 2007 when they were evaluating it, but they point out it is hard to determine which method it used to analyze its data. According to Ben-Simon and Bennett, Intellimetric developed in the late 1990s and was based "on a brain/mind model of information process and understanding" (p. 7). They further state that this system is based more on "artificial intelligence rather than on theoretical models of writing" (p. 7). Also developed in the '90s was the Intelligent Essay Assessor (IEA) which evaluates the content of the essay, mainly matching the student essays with other essays of similar score level. According to Ben-Simon and Bennett, IEA is a hybrid method because “content is arguably a key factor in evaluating writing quality” (p. 8). E-rater was also developed in the late ’90s and uses the brute-empirical method. However, the later version in 2003 was more “intuitively related to the characteristics of good writing” (p. 9) as Ben-Simon and Bennett observe. They mention e-rater judges a student’s writing with respect to “grammar, usage, mechanics and style; organization and development; topical analysis; word complexity and essay length” (p. 9).

Though some of these four writing systems of AES link computer-generated features to characteristics of good writing, these approaches, as Ben-Simon and Bennett (2007) observe, “do not explicitly link to specific features of writing attributes embedded in the rubrics for a particular testing program” (p. 13). This is because developers intend their automated systems to be general enough for a wide range of assignments.

Advantages of Using AES

According to Attali and Burstein (2006), researchers who investigated the e-rater v.2, which is a program used by the Educational Testing Service (ETS), the advantages of AES are that they use objective rating scales and allow greater standardization of scoring, “specifically allowing a single scoring model to be developed for all prompts of a program or assessment” (p. 23). AES systems are also “designed to find optimal solutions with respect to some measure of agreement between human and machine scoring” (Attali & Burstein, 2006, p. 13). They point out that e-rater also has “perfect inter-rater reliability” (p. 21). Notwithstanding these advantages, they say an AES system is not easily explained to users, which is “a threat to the face validity of AES” (p. 13), and it also has different solutions in different applications. As can be seen, AES therefore has enormous potential but definitely comes at a cost.

Computers have many distinct advantages, as Goldberg, Russell, and Cook (2003) found: “Computers seem to motivate students,” especially those who are “reluctant writers” (p. 18). Goldberg, Russell and Cook also discovered that “when using computers students also tend to make revisions while producing rather than after producing texts” (p. 20). They say that students “tend to make more revisions to their work especially if they use computers" (p. 20). Their most surprising finding was that "students who develop their writing skills while using the computer produce written work that is .4 standard deviations higher in quality than those students who write on paper" (p. 19). Keyboard skills also had an impact on writing. Goldberg, Russell, and Cook (2003) found that when students improved their keyboard skills, "the amount of time required to produce writing on computers would decrease" (p. 18). When students had learned to work efficiently with the keyboard, they could type automatically and therefore concentrate more fully on the quality of their writing rather than be bothered about handwriting and legibility.

Just the thought that a computer can score papers without consuming teacher time is a welcome thought today. Other advantages of using computer-based assessment would be the saving of teaching time, as Chalhoub-Deville (1999) observes: "Computerization accelerates the placement operation and reduces the human resources that are needed for the administration of the test, the marking and the production of student lists" (p. 122). However, as Dodigovic (2005) warns, "we need utmost caution in deciding on what is wanted from these innovations" (p. 105). She goes on to mention that, though automated essay grading will help in assessment and is on its way into standard practices because it "has clear efficacy advantages over human rating . . . it also has a potential to redefine the learning activities" (p. 105).

Other placement issues that are negated when students use a computer are handwriting and subjectivity issues. Harmer (2001) comments, "a computer screen frequently allows students to see their writing more objectively" and it also "removes the problem of bad handwriting" (p. 261). Some students these days may be more comfortable working on a personal computer than they are with pen and paper, so using a computer for placement tests will help in solving such placement issues.

Activities in the classroom involving computers can be asynchronic, meaning that different users can log onto the computers at different times to do the task like a discussion board activity. A synchronic activity is one which would require all students to be on their computers at the same time, for example an online class, or a chat room devoted to the class (Dodigovic, 2005).

Today large examination boards opt for computer scoring to combat administrative problems, and there are many products “commercially available to a multiplicity of users” (Dodigovic, 2005, p. 105). For example, testing bodies create tests which can be purchased by institutions for their own test use.

Criterion^{SM}

Criterion^{SM} , the e-rater technology being used in this research, “is a web-based service which evaluates a student’s writing skill and provides instantaneous score reporting and diagnostic feedback” to both writing instructors and students, as mentioned by Attali and Burstein (2006, p. 7), and will give a holistic score ranging from 1 to 6 and an analytical score based on 1) grammar, usage and mechanics; 2) style; and 3) organization and development. According to Attali (2004), an ETS researcher, Criterion^{SM} is powered by the e-rater automated scoring technology developed by ETS. As Attali explains, e-rater scoring is an application of Natural Language Processing (NLP), a field of computer technology that has used computational methods to analyze characteristics of text for the past fifty years, and e-rater compares the student sample essay to other faculty-scored essays on its database and gives a score in relation to those essays.

The e-rater engine provides score reporting, according to Attali and Burstein (2006): “The diagnostic feedback is based on a suite of programs (writing analysis tools) that identify the essay’s discourse structure, it also recognizes undesirable stylistic features and evaluates and provides feedback on errors in grammar, usage and mechanics” (pp. 7–8). Attali and Burstein go on to elaborate on e-rater’s special features: “The writing analysis tool identifies five main types of grammar usage and mechanical errors, verb formation errors, wrong word use, missing pronunciation, and typographical errors” (p. 8). They say that e-rater detects all these violations statistically and in relation to its corpus.

According to Attali and Burstein (2006), “the system is trained on a large corpus of edited text, from which it extracts and counts sequences of adjacent word and part-of-speech pairs called bigrams” (p. 8). These bigrams are compared to “the level of their frequency with regard to the general corpus” (p. 8). Attali (2004) says that the strength of accuracy depends on the essays on e-rater's database. This factor is why the Educational Testing Service (2007, Electronic References section: Criterion Details) suggests that the online evaluation should not only be the sole measure of assessing student writing for college placement, and the Criterion^{SM} score combined with that of a faculty reader is considered a proper indicator of student performance.

Correlating Test Results

For the purposes of comparing the analytical EPT with the holistic EPT and Criterion^{SM} scored EPT, I plan to correlate these tests and investigate their correlation so as to evaluate the appropriateness of using the analytical scale as an additional measure. At no point does this study wish to prove equivalency of these three measures. In mathematics when we use the term "equivalency," we mean to state that when two measures are being evaluated they are said to be the same or have the same sets and are equal to each other in whatever way they are compared. In language testing, too, equivalency is possible when two variables are considered to be equal. As Davies et al. (1999) explain, "test equivalence is the relationship between two or more forms of the same test. Test forms may be equated or equivalent" (p. 198). They state that "equivalent forms of tests are constructed from the same test specifications in order to measure the same skills" (p. 198). However, in this research only a correlation between different forms of scoring is sought. Davies et al. (1999) define correlation as "a procedure which measures the strength of the relationship between two or more sets of measures which are thought to be related" (p. 35).

For the purpose of correlating test scores, statistical measures are used to provide *a posteriori* evidence, such as scatterplots and histograms, the Pearson correlation coefficient to determine the strength of correlation between the two tests, and overlapping variance to indicate to what extent the respective two measures are evaluating the same skill. Cronbach's alpha evaluates inter-rater reliability (Davies et al.1999). According to Bailey (1998), in the scatterplot, the two axes represent the two variables and determine if there is a patterned linear relationship between two measures. The scatterplot provides a pictorial representation of the relationship between the two variables, and only if a linear relationship exists between the two measures can the

Pearson coefficient be used. Also, the Pearson requires normal distribution of scores, which is indicated by a histogram.

The Pearson correlation coefficient, also known as Pearson's "r," is the most well known and "widely used member of the correlation family in language assessment" (Bailey, 1998, p. 116). Using this measure provides a value showing the degree to which two variables are related (McNamara, 2000). Both sets of numbers have to be interval or ratio scales, where numbers within each set are independent of one another. Both distributions need to be symmetrical, and the numbers should have a linear relationship between the two sets, and a normal distribution which will be easily visible on the scatterplot (Bailey, 1998).

According to Bailey (1998), a Pearson's correlation coefficient is between −1.00 and 1.00, and "the closer the value is to the whole number [+ or – 1.00] the stronger the relationship between the two variables" (p. 113). Bailey says that values from 0.85–0.99 can be considered evidence of strong correlation, correlation coefficients of 0.70–0.84 can be considered moderately strong, and those between 0.45–0.69 can be termed moderate. She says that values below 0.45 may be considered "weak to moderate" (p. 113). She also points out that a positive correlation indicates that as scores on one variable increase they would reflect on the increase of the other variables, too. A negative correlation would demonstrate that if the scores of one variable decrease then scores on the other variable will increase. Thus the Pearson's correlation coefficient is often used to help establish the validity of a test (Bailey, 1998), correlating a new test with an established measure in criterion-related evidence of validity.

When we correlate the results of two tests, "we wish to see to what extent they are measuring the same trait" (Bailey, 1998, p. 117). Bailey says that we use the "r" (Pearson correlation coefficient) to correlate the students' scores on two measures, and we can square the resulting "r" values to find an overlapping variance between the respective two tests. According to Bowen, Madsen, and Hilferty (1985), "squaring your correlation is an important way to check your test's validity; that is whether it is in fact measuring what it purports to measure" (p. 386). Thus, the reason why overlapping variance needs to be calculated and tells us more than correlating statistics is that it interprets, as Bailey (1998) points out, "the extent to which the two tests being correlated measure the same construct. . . . sometimes called 'shared variance' or 'shared overlap'(p. 118). She explains that the stronger the magnitude of the correlation the greater the overlapping variance will be" (p. 118). Bachman (2005) states that shared variance can also be termed "a coefficient of determination" (p.103) which is a "square of the correlation coefficient." Bachman (2005) gives an example of developing a new writing test to illustrate the validity of using shared variance in correlating two tests. He explains that even though the new test correlates with the older test by 0.58 which appears to be moderate correlation, when we square this correlation, the percentage of shared variance is only about 34 per cent. Even though this indicates only a low proportion of shared variance, Bachman (2005) says that "both tests can be used to provide complementary information, and thus [we can] decide to use both for grading purposes" (p.104).

Cronbach's alpha is "a measure of internal inconsistency and reliability" among raters, according to Davies et al. (1999, p. 39), who state that "alpha indicates how well a group of items measure the trait of interest by estimating the proportion of test variance due to common factors among the items" (p. 39). This measure basically looks at inter-rater reliability and calculates the rate of variance from 0 to 1.0. Davies et al. (1999) say that the alpha can indicate whether or not there is any dichotomy, or in other words, difference, in the scoring. Obviously, the higher the covariance, the higher the reliability will be. This is an important statistic to check inter-rater reliability. If there are any differences between how raters have marked certain points in the test, and they are not in agreement, this will result in the rate of covariance being lower, indicating that there is a difference between what both raters see as appropriate.

To sum up, it is here that we realize that validating a placement test is a complex issue involving *a priori* and *a posteriori* evidence and accurate placement is best accomplished with multiple parameters to indicate the level of student placements. Once students have been placed into classes, general classroom observation and in-class assessment during the first week prove crucial in identifying misplacements. As Freeman (1998) notes, "closely watching and noting classroom events as a participant" (p. 94) can help in identifying student misplacements to a great extent. Thus in order to address the problem of misplacements, we need to have more than one measure of assessment in placement testing.

METHODOLOGY

Multiple placement assessment measures are desirable in order to ensure valid and reliable student placement into the required courses, according to Coombe and Hubley (2003) and Norris, Brown, Hudson and Yoshioka (1998). The EPT at AUS does successfully place the majority of students into appropriate levels of WRI 001, WRI 101 and WRI 102, but there is a small percentage of students whom teachers feel are "borderline" and/or students feel are "misplaced."

To address the issue of misplacements I decided to undertake a study into the EPT at AUS. My research questions were thus:1) What does *a priori* and *a posteriori* evidence reveal about the validity and the reliability of the proposed analytical scale for use with the EPT in the Fall 2007 semester? 2) Who, if any, are the "borderline/potentially misplaced" students in WRI courses for the Spring 2008 semester? 3) In addition to the holistic EPT, what do the analytical EPT results and Criterion^{SM} reveal about the appropriacy of placement of the identified "borderline/potentially misplaced" students? My original third research question was to use the holistic EPT, the analytical EPT and the in-class writing samples to re-evaluate the placements of the identified " borderline/potentially misplaced " students in Spring 2008. The third research question was revised to add Criterion^{SM}, as the in-class writing samples were not available. Criterion^{SM} was used instead of the in-class writing samples (because it was readily available although involving a subscription fee) to provide additional evidence for the issue of placements for the identified potentially misplaced students in Spring 2008.

I selected a representative sample of 65 EPT scripts which were individually scored by Rater 1 and Rater 2 analytically using the Jacobs et al. (1981) analytical scale which, according to Hamp-Lyons (1995b), is a widely used scale in North American universities today. After analytically scoring the EPT, the scripts were typed by Rater 1 and Rater 2 with all their errors, and then these scripts were input into Criterion^{SM}, which gave a holistic score between a range of 1–6. These scores were then converted to z scores and later to T-scale scores to investigate the strength of their correlation through Pearson's " r."

Participants

The participants consisted of two groups: 1) 65 students representatively selected from WRI 001, WRI 101, and WRI 102 who had taken the EPT in Fall 2007, and 2) 44 WRI students who had taken the EPT in the Spring of 2008 and who had been considered “borderline/potentially misplaced” by DWS teachers. All these students were English-as-a-second-language speakers who had varying backgrounds depending on when English was introduced in their K-12 curriculum. The students were Arabs from the Middle East, or Asians from Pakistan, India, Sri Lanka and the Philippines, and had different proficiency levels in English, reflected in their WRI placement levels. The first languages of these participants were Arabic, Urdu, Hindi, Tagalog, or Sinhalese. Thus, the participants of this research fall into two groups: 1) a representative sampling of all newly admitted WRI students (Fall 2007) and 2) a group identified by their teachers as being “borderline/potentially misplaced” (Spring 2008). Out of the 65 students in Fall 2007, 32 students were from WRI 001, 14 students were from WRI 101, and 19 students were from WRI 102. In Spring 2008, out of the 44 students whom DWS teachers identified as “borderline/potentially misplaced,” there were 5 students identified from WRI 101 and 39 students identified from WRI 001.

Data Collection

Fall 2007

Data were collected in the Fall of 2007 in order to validate the use of the Jacobs et al. (1981) analytical scale in the context of the EPT at AUS. This data consisted of 65 scripts of the students who had already been given their EPT in the Fall of 2007. The total of 600 EPT scripts of Fall 2007 were kept in the Department of Writing Studies office. Every fifth paper from all of the Fall 2007 scripts was taken as I wanted to ensure that all teachers who had marked the EPT were included in the sample, but this sampling of all teachers’ scoring could not be achieved as the papers were arranged according to student numbers, instead of by the teachers scoring them. Thus, I could not ascertain whether or not these scripts selected were an apt cross section of the EPT scored by the DWS faculty raters. Though 100 scripts were selected, some of these students could not be located on the DWS roster, and I had to keep adding more scripts (every fifth paper) verifying enrollment three times. This time-consuming process resulted in selecting 97 scripts. Finally, I had 35 scripts of WRI 001, 38 scripts of WRI 101, and 24 scripts of WRI 102 (a total of 97, all which were identified as those of students who had taken the Fall 2007 EPT and were attending class that semester). I gave out 97 consent forms to the teachers concerned and received 65 signed consent forms, because 22 students did not consent to participate in the research and 10 forms could not be accounted for. Thus I was left with 65 scripts: 32 scripts from WRI 001, 14 scripts from WRI 101, and 19 scripts from WRI 102, representing all three WRI levels. These handwritten EPT scripts were double scored analytically by Rater 1(myself) and Rater 2 (an editor of a publishing house) and then transcribed by both of us and saved as Microsoft Word documents which were double checked by each for the other to determine accuracy.

Spring 2008

Data were also collected in the Spring of 2008. The DWS teachers who taught all three levels were asked to identify students whom they felt were borderline. There were 116 students from a total of 1407 students registered for Spring 2008 identified by seven DWS faculty as being “borderline/potentially misplaced.” After obtaining student consents, the next step was to collect these students’ EPT scripts from the DWS office. On searching for these scripts, it was found that EPTs were not available for 53 of these students because they came through WRI 001 and WRI 101 and were not placed directly into those classes through the Spring 2008 EPT. Out of the remaining 63 scripts, 16 students’ EPT scripts could not be found in the Spring 2008 pile, which meant that these students would have taken an earlier EPT. Out of the 47 scripts remaining, three scripts (which were administered at three separate times) could not be considered for this research as the reading prompts for these papers were different from the rest of the papers which had either Prompt 1 (21 students) or Prompt 2 (23 students). (Both these prompts can be found on pages 40 & 41.) However, these three scripts were used to practice uploading documents to Criterion software before inputting the 44 participating students’ scripts for analysis. The 47 scripts were analytically double scored, and the other statistical analyses described earlier were carried out with these scripts.

Answering the Research Questions

For an in-depth understanding of the rationale behind the methodology, we need to look at how the research questions were addressed.

Research Question 1: What does *a priori* and *a posteriori* evidence reveal about the validity of the proposed analytical scale for use with the EPT in the Fall 2007 semester?

A Priori Evidence

A priori evidence is evidence that is required before the actual research is carried out. In the context of this research, the use of the analytical scale by Jacobs et al. (1981) was proposed to investigate misplacements in the EPT. *A priori* evidence was necessary to validate this scale, and this was done in Fall 2007 before the placement evaluation research was carried out in Spring 2008. The analytical scale by Jacobs et al. (1981) was chosen for use in this research as it is a scale widely used in North American universities. In order to choose an appropriate analytical scale to score the EPT, I examined the test specifications of the holistic EPT, together with the "prompt attributes" and "response attributes" (Lloyd, Davidson & Coombe, 2005, pp. 18–19) to see if the same construct was addressed by these two measures, the holistic EPT rubric and Jacobs et al. scale. The prompt attributes as described by Lloyd, Davidson and Coombe (2005) are the "directions or instructions that the test taker will read" (p. 18), while the response attributes will specify details as to "how the test taker will respond to the item or task" (pp. 18–19).The writing prompts of the EPT (see Appendix A. Sample of the English Placement Test) required the students to write an essay in paragraphs about a topic which was related to a provided reading prompt. (This essay was usually about topics which were common knowledge to high school students, and needed no specific technical knowledge to be answered.)

The prompts for Fall 2007 and Spring 2008 are listed below for reference.

Fall 2007

Prompt 1: Consider two cultures you are familiar with. How do communication styles in these cultures differ? Be specific in your analysis.

Prompt 2: Wong describes the clash of two cultures and the conflicts that can occur from it. Do you think it is possible for someone to maintain connections with his or her original culture and at the same time become an “all American”? What does one gain or lose in becoming completely. (Wong refers to the article included in the reading prompt.)

Spring 2008

Prompt 1: Write an essay in which you explore your ideas about the cultural complexities young students face upon entering a multicultural university like the American University of Sharjah. What skills do you believe are necessary for students to have in order to function well within this culturally diverse community? Be sure to provide details and examples to support your ideas.

Prompt 2: Consider the various types of technology at use now – cellphones, Ipods, Internet – and formulate your own position on how/whether they have changed people’s behavior. Then, in a well-developed essay, offer your point of view on technology’s impact on social interaction. Be sure to provide details and examples to support your ideas.

The response attributes (Lloyd, Davidson & Coombe, 2005) were for students to write an essay in several well developed paragraphs about a given topic. (See Appendix A. Sample of the English Placement Test.) This test was an hour-long pen/pencil and paper test. The language used by the test takers was to be formal, and the students were required to express a point of view. The structural requirements were that the essay needed to include an introduction, a thesis statement, an appropriate number of paragraphs and a conclusion. Content, coherence, clarity of expression, grammar, punctuation and readability were taken into consideration when assessing the essays.

Students were asked to leave ample time for proofreading their essays. The formatting requirements included leaving a blank line between written lines to ensure that students had enough space to put in changes if necessary and also to increase the readability of the essay. Items like organization, style, coherence, unity and vocabulary were looked at when the EPT raters scored these papers holistically.

In the analytical profile by Jacobs et al. (1981), the underlying construct being evaluated is the same as the holistic EPT. In the profile, the five component scale consists of content, organization, vocabulary, language use and mechanics. Jacobs et al. (1981) also state that "the most reliable indicator of a writer's ability (as demonstrated on any particular writing task) comes from a sum of the five judgments, rather than from individual component scores" (p. 32). Thus, Jacobs et al. are of the opinion that these five component scales should not be considered independently but rather as a sum of a student's performance in writing, much like the holistic evaluation of the EPT at AUS. This was why the analytical profile was considered apt, in *a priori* evidence for this research.

The holistic EPT papers had been double marked by DWS faculty on a 6-point scale and rated by a third rater if there was more than a point difference between the two raters. Previously, each semester this task was administered by DWS, but in Fall 2007 and Spring 2008 students wrote their EPT at different times and with different reading prompts as the EPT was conducted by AUS at the Testing Center those semesters.

A Posteriori Evidence

It was important to demonstrate that the analytical scale was a reliable and valid measure for evaluating the EPT, and it is for this reason that *a posteriori* evidence was necessary before the study on misplaced students was conducted in Spring 2008. Once the 65 students' scripts of Fall 2007 were obtained with the consent of the students, these 65 scripts were double scored by Rater 1 and Rater 2 using the analytical scale. (See Appendix C for a copy of the Jacobs et al. analytical scale.) Marking each of these 65 scripts differently using the analytical scale proved to be a challenging task as both raters had to become familiar with this analytical scale. It took a week of two-hour sessions to understand how to mark scripts according to this scale (informal inter-rater training/calibration). Both Rater 1 and Rater 2 read the ESL Composition Profile by Jacobs et al. (1981) which clearly explained how to mark these scripts. After double marking these scripts, the two raters typed the scripts, turning off the spell check on their computers so as to have authentic printed representations of the students' essays. Once the essays were typed, they were evaluated with Criterion^{SM} , resulting in a holistic score on a 6-point scale.

The analytical essays were on a 100-point scale, so to ensure comparability, T-scale scores were used. The conversion to T-scale scores was necessary as each of them had different scales of measurements: the holistic EPT was marked on a 12-point scale, the analytical EPT was marked on a 100-point scale, and Criterion^{SM} was marked with a 6-point scale. The T-scale score, according to Davies et al. (1999), "is a transformation of a z score, equivalent to it but with the advantage of avoiding negative values" (p. 194). Bailey (1998) offers a simpler explanation: "Multiplying the z score by 10 gets rid of a decimal place, and adding 50 gets rid of any minus signs" (p. 104).

First, the raw scores, or the scores as they were on paper, were converted to z scores. The z score as expressed by Davies et al. (1999), "is a way of placing an individual score in the whole distribution of scores on a test; it expresses how many standard deviations units lie above or below the mean" (p. 228). They state that "scores above the mean" are considered "positive" and "scores below the mean" are considered "negative" (p. 228). This conversion to z scores was done using the SPSS software. These z scores were converted to T-scale scores using Microsoft Excel. An example of the T-scale scores can be seen in Table 6 and 7 (Results of Fall 2007 for Prompt 1 and Results of Fall 2007 for Prompt 2) in the following chapter on page 70 and page 75, respectively.

To provide *a posteriori* evidence of the validity of the analytical scale for use with the EPT as an additional measure for identifying misplaced students, statistical analysis was carried out using the converted scores. Cronbach's alpha was calculated in order to evaluate inter-rater reliability.

Cronbach's alpha is, according to Davies et al. (1999), "a measure of internal inconsistency and reliability" among raters (p. 39). Cronbach's alpha was used to examine the extent to which both raters were similar in their marking. In other words the inter-rater reliability of the analytical scored EPT scripts was calculated by the variance from 0 to1.0. Davies et al. (1999) say that the alpha could indicate whether or not there are any differences in the scoring pattern. Obviously, the higher the covariance between the two raters, the higher the reliability would be. Cronbach's alpha was an important statistic to check inter-rater reliability, when the analytical EPT was double scored.

Once a high inter- and intra-rater reliability was determined, a combined analytical score was used as well. Correlations were carried out between the holistic EPT and the Analytical EPT Rater 1, the holistic EPT and Analytical EPT Rater 2, and the holistic EPT and the combined analytical EPT separately. A combined analytical score would be of practical use in the classroom of borderline/potentially misplaced students as it would provide stronger evidence for the use of this additional measure.

Only if there was high inter-rater reliability could the Pearson correlation coefficient be used. Using Pearson requires a linear relationship between the variables (test scores) and a normal distribution, which are indicated by scatterplots and histograms. According to Bailey (1998), in the scatterplot, the two sets of axes represented the combinations of two variables (holistic EPT and analytical EPT, Criterion^{SM} and analytical EPT) and would indicate if there was a patterned linear relationship between the two measures. The scatterplot would provide a pictorial representation of the relationship between the two variables, and the histogram would indicate the distribution of scores.

The measure used to determine the strength of correlation was the Pearson correlation coefficient, also known as Pearson's "r." The reason why this particular statistic served my purpose was because by using this measure I could arrive at a value showing the degree to which two variables (e.g., holistic EPT and the analytical EPT, Criterion^{SM} and the analytical EPT) were related (McNamara, 2000). Both sets of numbers had to be interval or ratio scales, which the analytical, holistic, and Criterion^{SM} scores were, where numbers within each set were independent of one another. Both distributions needed to be symmetrical, and the numbers needed to have a linear relationship and a normal distribution between the two sets, which would be easily visible on the scatterplot and the histogram.

An important application of correlation statistics in this research was to use the analytical EPT scores and correlate them with the scores of the established holistic EPT and Criterion^{SM}. The resulting Pearson correlation coefficients were considered to be an estimate of the new analytical scale's criterion-referenced evidence of validity where the accepted measures (holistic EPT and Criterion^{SM}) were taken as the criteria. In this research the analytical EPT scores were correlated with the holistic EPT scores and the Criterion^{SM} scores using the Pearson correlation coefficient to determine the strength of correlation between the holistic and the analytical EPT scores, and between the Criterion^{SM} and the analytical EPT scores.

When correlating the results of two tests, it was necessary "to see to what extent both tests are measuring the same trait" (Bailey, 1998, p. 117). Bailey says that we can square the resulting "r" values to find an overlapping variance between two tests. According to Bowen, Madsen and Hilferty (1985), squaring the correlation is an important way to check the test's validity by finding out whether the test is actually measuring what it seeks to measure. Thus, the reason why the overlapping variance needs to be calculated, as Bailey (1998) points out, is to find out "the extent to which the two tests being correlated measure the same construct," sometimes termed as "shared variance" or "shared overlap" (p. 117). The stronger the correlation, the greater the overlapping variance.

Research Question 2: Who, if any, are the borderline/potentially misplaced students in WRI courses in the Spring 2008 semester?

After having established a correlation between the analytical EPT and the holistic EPT/ Criterion^{SM}, it then became possible to conduct the potential misplacement study in the spring semester of 2008. It was important to identify "borderline/potentially misplaced" students, and this was meant to be done through classroom observation by DWS teachers and an in-class writing assignment. Within the first week of the semester, all DWS writing teachers gave their students in-class writing assignments in order to assess the students' level of English language proficiency. (This was typically a thirty-minute writing task which was usually descriptive and differed from teacher to teacher so as not to pressure the students while they wrote.) This task was scored holistically, and teachers used it to identify “borderline/potentially misplaced” students. The DWS teachers were informally asked to identify “borderline/potentially misplaced” students in their classes, which they did. These 44 students were then asked for their consented use of their EPTs in this research.

Revised Research Question 3: In addition to the holistic EPT, what do the analytical EPT results, and Criterion^{SM} reveal about the appropriacy of placement of the identified “borderline/potentially misplaced” students? (Criterion^{SM} was used instead of the in-class writing samples, as they could not be obtained.)

In Spring 2008 all the teachers with the exception of one did not give the in-class writing assignments to me for further examination. Possible explanations could be that they were just too busy with other work, this was another piece of work to add to their workload, or they had different types of in-class assessments, or they plain forgot to get back to me. So I was not able to use the in-class writing assessment to evaluate the students’ Spring 2008 EPT performance. Only the holistic EPT, and the analytical EPT scores could be used. Thus, Criterion^{SM} was used as a third measure as it was readily available and already paid for by virtue of the AUS Seed Grant. (The Seed Grant was obtained by Dr. Betty Lanteigne for both my research and her research into the use of Criterion^{SM} at AUS.)

After they were analytically double scored, the Spring 2008 EPT scripts of the identified “borderline/potentially misplaced” students were accurately typed with the students’ original errors replicated in the typed versions. All 44 scripts were double checked to ensure reliability and validity in that the typed scripts were copies of the handwritten ones of the students. After all these papers were typed and saved as Microsoft Word documents, they were input into Criterion^{SM}, which scored them. When using Criterion^{SM} to grade student essays, I noticed that the time allotted by ETS for instructor-generated topics had to correspond to the standard higher educational requirement which was thirty minutes. This was contrary to the one hour that we allot to the AUS EPT. This dichotomy in time was resolved by resetting the Criterion^{SM} AUS EPT as thirty minutes in the Criterion^{SM} settings. The different timings were deemed appropriate because the test tasks differed. The Criterion^{SM} prompt is a one/two sentence description of a persuasive topic whereas the EPT task has a reading prompt. This did not alter the results of this study because this study does not aim at equivalency between the Criterion^{SM} task and the EPT task. It only correlates the results of both. Tasks overlapped but were not identical, but the output evaluated was the same: a persuasive essay evaluated by similar criteria based on the TWE.

Thus the last stage of this research involved correlating the holistic scores and the analytical EPT scores, in order to reevaluate the case of "borderline/potentially misplaced" students. The in-class writing sample could not be collected as teachers did not hand in copies of the in-class writing sample. This was indeed unfortunate as the in-class writing samples would have added another measure to evaluate student misplacements in Spring 2008. Therefore Criterion scores were used instead of in-class assignment scores and served as a complementary measure in examining the discrepancies between the holistic and the analytical EPT scores, since Criterion measured the same features addressed by the holistic and analytical EPT. In this triangulated picture of Spring 2008 student writing, T-scale scores, as suggested by Bailey (1998), were used to give a common basis of comparison because each measure used a different point scale.

FINDINGS

In order to analyze the findings of this research, the research questions need to be revisited. The first research question sought to investigate *a priori* and *a posteriori* evidence about the validity and the reliability of the proposed analytical scale for use with the EPT in the Fall 2007 semester. The second research question was concerned with the identification, if any, of borderline students in the WRI courses for the spring 2008 semester. Finally, the third research question dealt with the results of the analytically scored EPT in comparison with the holistically scored EPT and Criterion^{SM} and what it revealed about the appropriacy of placement of these identified "borderline/potentially misplaced" students.

Research Question 1:

To provide *a priori* evidence of validity, the analytical scale used to score the EPT as an additional measure needed to be an analytical scale which is widely accepted. A scale which meets this criterion is the ESL (English as a second language) Composite Profile by Jacobs et al. (1981). This scale has five weighted components with content the first, which is most heavily rated with 30 points; the other components are organization rated as 20 points, vocabulary rated as 20 points, language use rated as 25 points, and mechanics rated as 5 points. (Refer to Appendix D for example of Analytical Scale.) Each of these components is further subdivided into "four masterly levels: excellent to very good, good to average, fair to poor and very poor" (Jacobs et al., 1981, p. 31). The reason I chose this analytical scale is because it included similar evaluation criteria to those used by the DWS faculty. Even while holistically marking the EPT, DWS raters considered content, organization, vocabulary, language use and mechanics on a scale of 1 to 6, similar to the 6-point scale that Test of Written English follows, as mentioned in Table 1. (See Appendix C for EPT scoring guidelines.)This analytical scale also has similar evaluation criteria to those of Criterion^{SM}, the online assessment developed by ETS to evaluate student essays which gives a holistic score ranging from 1 to 6 and trait feedback based on 1) grammar, usage and mechanics; 2) style; and 3) organization and development (can be found on Criterion website under Trait Analysis feedback menu). Criterion^{SM} has different levels of evaluation (from elementary to college), and the level used in this research was first year college, since the students are being evaluated by the EPT for placement into freshmen level writing courses.

Table 1

6-Point Scale as followed by TWE, Holistic EPT, and Criterion^{SM}

	TWE	Criterion^{SM}	Holistic EPT
6	Excellent	Excellent	WRI 102
5	Very good	Skillful	WRI 102
4	Good	Sufficient	WRI 101
3	Adequate	Uneven	WRI 101
2	Weak	Insufficient	WRI 001
1	Unsatisfactory	Unsatisfactory	WRI 001

To provide *a posteriori* evidence of validity of the analytical scale for use with the EPT as an additional measure for identifying misplaced students, different statistical analyses were carried out. In Fall 2007 the first statistic to be carried out was Cronbach's alpha to ascertain inter-rater reliability. Once I was certain that the inter-rater reliability was strong, I combined the scores of the two raters and established a combined analytical EPT score, after which I developed a series of scatterplots between the holistic and analytical EPT scores, and the Criterion^{SM} and analytical EPT scores, because to calculate the Pearson's correlation coefficient, demonstrating a linear relationship in the scatterplots was mandatory. Also, frequency of distribution of scores indicating normal distribution in the histograms meant I could go ahead and calculate the Pearson's correlation coefficient and the overlapping variance of the scores.

Results for Fall 2007

For the Fall 2007 EPT the prompts were as follows:

> Prompt 1: Consider two cultures you are familiar with. How do communication styles in these cultures differ? Be specific in your analysis.
>
> Prompt 2: Wong describes the clash of two cultures and the conflicts that can occur from it. Do you think it is possible for someone to maintain connections with his or her original culture and at the same time become an "all American"? What does one gain or lose in becoming completely Americanized?

Results for Prompt 1

According to Table 2 below, the Cronbach's alpha between Rater 1 and Rater 2 was 0.971, indicating a very high degree of inter-rater reliability.

Table 2

Cronbach's Alpha for Prompt 1, Fall 2007 between Rater 1 and Rater 2

Cronbach's alpha	Number of scripts
0.971	48

Scatterplots for Prompt 1 showed a linear relationship between the scores (refer to Appendix F, Scatterplots for Prompt 1 in Fall 2007).

Davies et al. (1999) define a normal distribution curve as a normal probability curve which shows "a similar symmetrical pattern or distribution. This distribution is bell-shaped with most [scores] being average, and extremes, both low and high being few" (p. 129). "In language tests," Davies et al. (1999) explain, "the distribution of scores in a population is [said to be] normally distributed when most test takers [score] around the average (or mean) and progressively fewer towards the extremes" (p.129). While the histograms for the holistic EPT and the analytical EPT showed a broad range of scores (see Figures 1 and 2), a normal distribution of scores was not apparent until the Criterion^{SM} histogram. (See Figure 3.)

Observing the histogram in Figure 1, we can see that eighteen students got a score of 4 and twelve students got a score of 10. The other students (numbering 15) were distributed in the middle with scores ranging from 5 to 9 and a few extreme scores. This shows a broad range of holistic EPT scores, as is visible in this histogram.

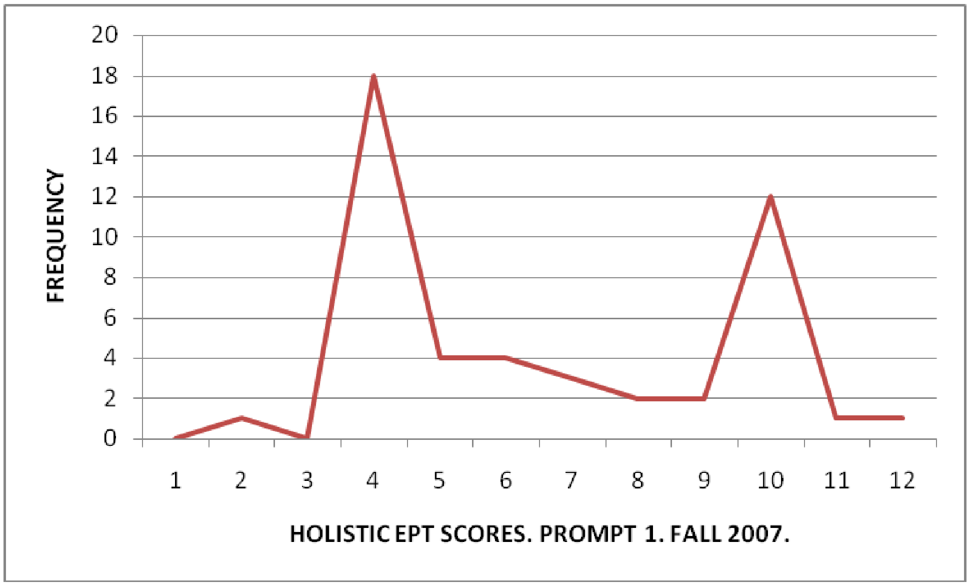


Figure 1. Histogram of Holistic EPT Scores for Prompt 1 in Fall 2007

n = 48

On observing the histogram in Figure 2, for the Analytical EPT we can find that twelve are in the middle range of 71 to 80, while twelve scores are also between 86 and 90 with sixteen of the scores being in the range of 51 to 70. Two scores were above the range of 86 to 90 and three scores were between 41 and 50.

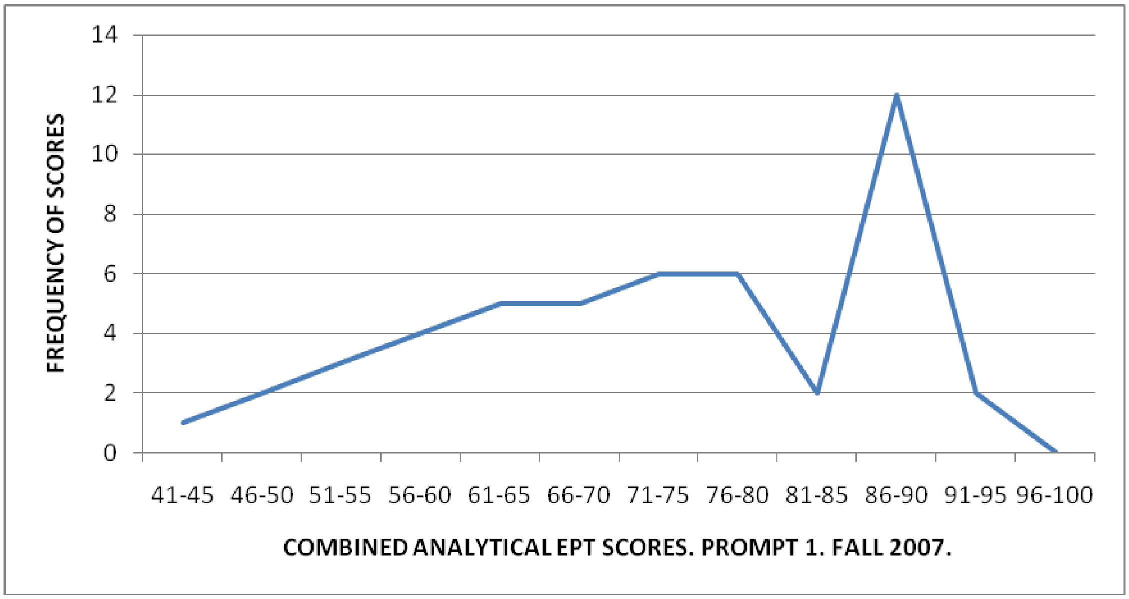


Figure 2. Histogram of Combined Analytical EPT Scores for Prompt 1 in Fall 2007

n = 48

Looking at Figure 3, which shows the distribution of combined analytical EPT scores for Prompt 1 in Fall 2007, we can at once see a broad range of distribution with fourteen students with a score of 3 and twenty students with a score of 4. There were ten scores of 5, and three scores of 2, but the majority of scores fell into the middle range as they should in a normal distribution. The two prior histograms of the holistic EPT and the combined analytical EPT both show a broad range of distribution of scores, but a normal bell curve is only visible in the Criterion^{SM} histogram.

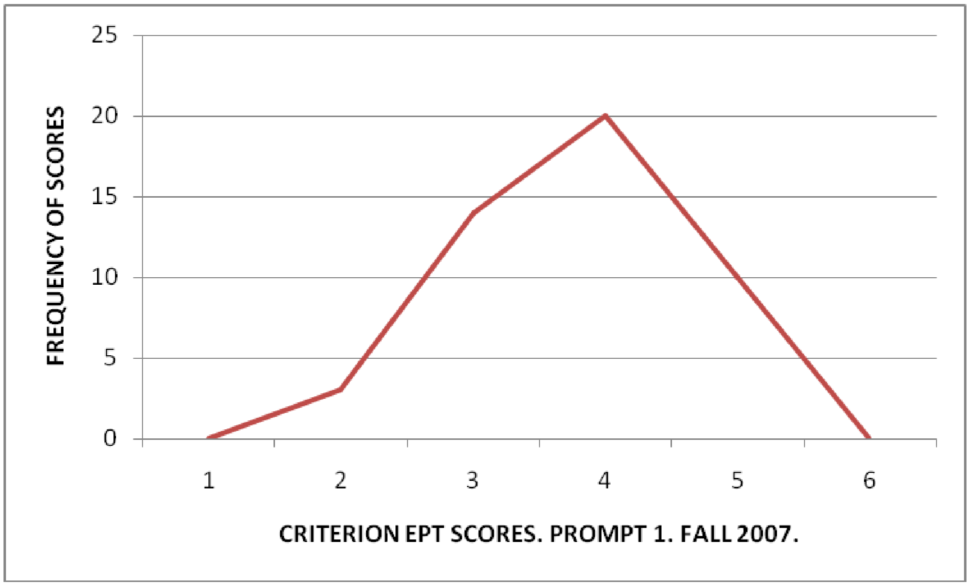


Figure 3. Histogram of Criterion^{SM} EPT Scores for Prompt 1 in Fall 2007

n = 48

Results of Prompt 2

Table 3 below indicates inter-rater reliability as shown by Cronbach's alpha for Prompt 2 showing a near perfect correlation between the two raters. Though the number of scripts (17) was smaller compared to Prompt 1 (48), this prompt had a higher inter-rater reliability between Rater 1 and Rater 2. One possible explanation is that these scripts were scored after Prompt 1, which meant that both raters had calibrated their marking by this time.

Table 3

Cronbach's Alpha for Prompt 2 , Fall 2007 between Rater 1 and Rater 2

Cronbach's alpha	Number of scripts
0.990	17

Figure 4 follows on the next page

Observing the holistic EPT histogram in Figure 4, we can find most of the students (seven each for both scores) scoring either 4 or 10. Only one student scored in the middle range of 5 to 9, and only two students scored less than 4. While histograms for the holistic EPT and the analytical EPT showed a broad range of scores (see Figures 4 and 5), a normal distribution of scores was not apparent until the Criterion^{SM} Histogram (Figure 6).

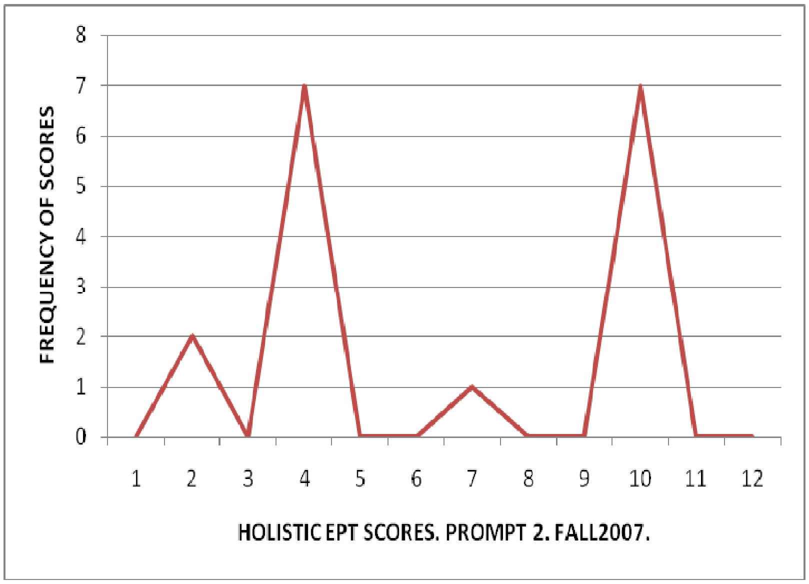


Figure 4. Histogram of Holistic EPT Scores for Prompt 2 in Fall 2007

n = 17

On observing the pattern in the histogram in Figure 5, we can see the greatest number of scores falling in two ranges: seven students' scores in the range of 46 to 65, and seven scores within the range of 86 to 95. There were only two students who scored in the middle range of 66 to 85. Though the scores are distributed in two groups at the ends of the histogram, it is still a broad range of score distribution.

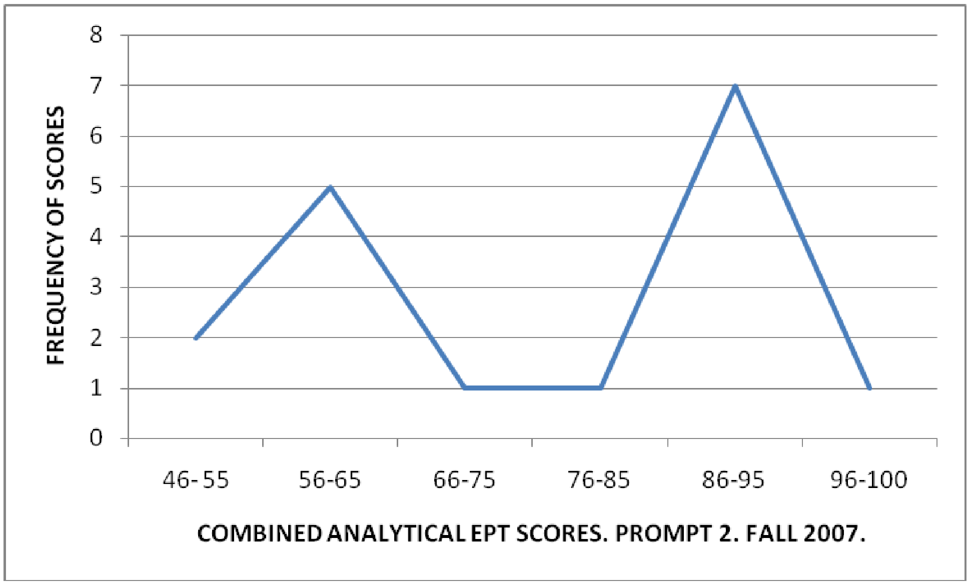


Figure 5. Histogram of Combined Analytical EPT Scores for Prompt 2 in Fall 2007

n = 17

Observing the histogram in Figure 6, we find the scores in the middle level of 3 having eight students and 4 having five students, suggesting again that most of the students were placed in this middle range. Three students got less than 3 as a score and two got 5 as a score, making a normal distribution. Though the histograms of Figure 4 and 5 showed a broad range in the distribution of scores for Prompt 2, a normal bell curve was only visible in the Criterion^{SM} histogram which is presented in Figure 6.

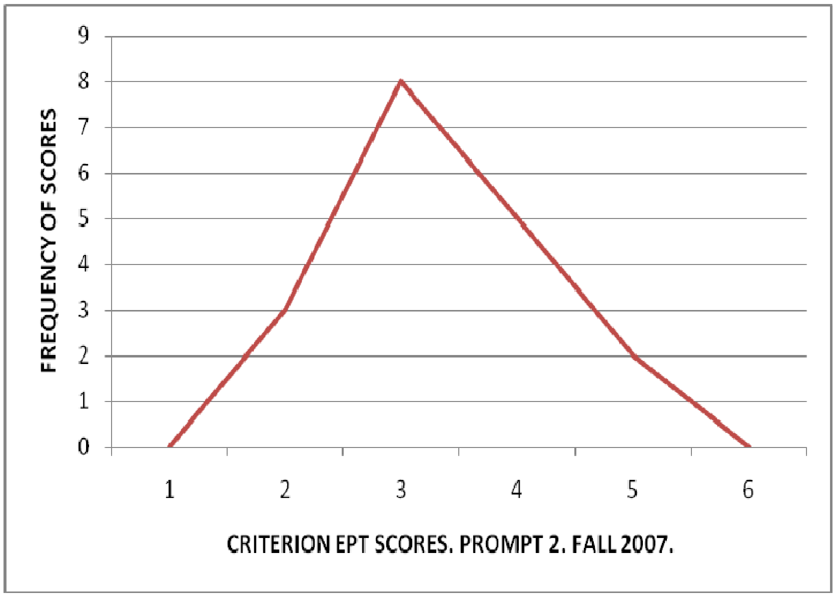


Figure 6. Histogram of Criterion^{SM} EPT Scores for Prompt 2 in Fall 2007

n = 17

The test takers' essay scores were normally distributed in the Criterion^{SM} scoring while the holistic and the analytical EPT scoring of these same essays was not. Spearman correlation coefficient is another correlation measure which requires ranked scores and is not as sensitive to normal distribution and sample size as Pearson correlation coefficient is. However, Spearman is affected by tied scores, of which there were many in these results. Therefore Pearson correlation coefficient was utilized to investigate the correlation between holistic and analytical scoring and analytical and Criterion^{SM} scores.

Results of Prompt 1: Pearson's Correlation Coefficient and Overlapping Variance

After analyzing the scatterplots and histograms for Prompt 1 and Prompt 2, the next step was to analyze the statistical correlations of the Pearson correlation coefficient and the overlapping variance between the three measures (holistic EPT, analytical EPT and Criterion^{SM}).

Table 4 shows that Pearson's correlation coefficient for Prompt 1 between the holistic and the analytical Rater 1 scores was moderately strong at 0.809, while between the holistic and Rater 2 scores there was a strong correlation at 0.875. The holistic EPT score in correlation with the combined analytical EPT score of Rater 1 and Rater 2 was 0.854, which implies that the correlation is strong. [As Bailey (1998) explains, "the closer the value is to the whole number [+ or – 1.00] the stronger the relationship between the two variables" (p. 113). Bailey says that values from 0.85 to 0.99 can be considered evidence of strong correlation. Correlation coefficients of 0.70 – 0.84 are considered moderately strong, and those between 0.45 – 0.69 are deemed moderate. Values below 0.45 are considered weak to moderate.] Observing Table 4 below, we can see that the analytical EPT is more strongly correlated to the holistic EPT than to Criterion^{SM} for Prompt 1. Figures show that the analytical combined score correlates more closely to the holistic EPT with a correlation coefficient of 0.854, whereas Criterion^{SM} only correlates with the combined score moderately with a coefficient of 0.663.

Table 4

Fall 2007. Results of Prompt 1. Holistic and Criterion^{SM} Scores Compared with Analytical scores

Fall 2007. Results of Prompt 1.						
	Holistic & Rater 1	Holistic & Rater 2	Holistic & Combined	Criterion & Rater 1	Criterion & Rater 2	Criterion & Combined
Pearson Correlation Coefficient	0.809 Moderately Strong	0.875 Strong	0.854 Strong	0.614 Moderate	0.695 Moderate	0.663 Moderate
Overlapping Variance	0.655	0.766	0.729	0.377	0.483	0.440

While comparing the Criterion^{SM} scores with the analytically scored EPT, it was found that the Pearson's correlation coefficient between Criterion^{SM} and Rater 1 was moderate at 0.614, and the Rater 2 score showed a higher moderate correlation at 0.695. The combined analytical scores ranged in between at 0.663, suggesting only a moderate correlation between Criterion^{SM} and the analytically marked EPT with Prompt 1.

The next step was to measure the overlapping variance between Criterion^{SM} and Raters 1 and 2. The overlapping variance is calculated by squaring the "r" values, as Bailey (1998) states, to know "to what extent both tests are measuring the same trait" (p. 117). Bowen, Madsen and Hilferty (1985) say that squaring the correlation is an important way to check the test's validity by finding out whether the test is actually measuring what it seeks to measure. Thus, the reason why the overlapping variance needs to be calculated, as Bailey (1998) points out, is to measure "the extent to which the two tests being correlated measure the same construct," sometimes termed as "shared variance" or "shared overlap" (p. 117). The stronger the correlation, the greater the overlapping variance.

When observing the overlapping variance between Criterion^{SM} and Rater 1 scores, the overlapping variance was calculated at 0.377 indicating 37.7% overlap, while the overlapping variance between Criterion^{SM} and Rater 2 was slightly higher at 0.483 with 48.3% overlap. The combined analytical score and Criterion^{SM} showed a shared variance of 0.440 (44% overlap). While observing the holistic and analytical scores and their overlapping variance, it was seen that the overlapping variance between Rater 1 and the holistic EPT was 0.655 (65.5% overlap), while the overlapping variance for Rater 2 and the holistic was higher at 0.766 (76.6% overlap). The combined analytical scores with the holistic scores showed an overlapping variance of 0.729 (72.9% overlap).

Results of Prompt 2: Pearson's Correlation Coefficient and Overlapping Variance

Table 5 shows that Pearson's correlation coefficient for Prompt 2 between the holistic and the analytical Rater 1 scores was 0.934, and between the holistic and Rater 2 it was 0.953. The holistic EPT score compared to the combined analytical EPT score of Rater 1 and Rater 2 was 0.949, which is evidence of strong correlation, according to Bailey (1998).

Table 5

Results of Prompt 2

Fall 2007. Results of Prompt 2. Holistic and Criterion^{SM} Scores Compared with Analytical Scores.

Fall 2007. Results of Prompt 2.						
	Holistic & Rater 1	Holistic & Rater 2	Holistic & Combined	Criterion & Rater 1	Criterion & Rater 2	Criterion & Combined
Pearson Correlation Coefficient	0.934 Strong	0.953 Strong	0.949 Strong	0.790 Moderately strong	0.812 Moderately strong	0.806 Moderately strong
Overlapping Variance	0.872	0.908	0.901	0.624	0.659	0.650

While comparing the Criterion^{SM} scores with the analytically scored EPT for Prompt 2, it was found that the Pearson's correlation coefficient between Criterion^{SM} and Rater 1 was 0.790, while the Criterion^{SM} and Rater 2 score was slightly higher at 0.812. The combined analytical scores correlation ranged in between them at 0.806, suggesting a moderately strong correlation between Criterion^{SM} and the analytically marked EPT, much stronger than the correlation for Prompt 1, which was a moderate 0.663.

When observing the overlapping variance between Criterion^{SM} and Rater 1 scores, the overlapping variance was calculated at 0.624 (62.4% overlap), while the overlapping variance between Criterion^{SM} and Rater 2 was slightly higher at 0.659 (65.9% overlap). The combined analytical score and Criterion^{SM} showed a shared variance of 0.650 (65.0% overlap). While observing the holistic and analytical scores and their overlapping variance, it was seen that the overlapping variance between Rater 1 and the holistic EPT was 0.872 (87.2% overlap), while the overlapping variance for Rater 2 and the holistic was higher at 0.908 (90.8% overlap) (which is by far the strongest covariance). The combined analytical scores with the holistic scores showed an overlapping variance of 0.901 (90.1% overlap).

My observations on the results centered mainly on the prompts and their effectiveness. Prompt 1 was "Consider two cultures you are familiar with. How do communication styles in these cultures differ?" and Prompt 2 was "Wong describes the clash of two cultures and the conflicts that can occur from it. Do you think it is possible for someone to maintain connections with his or her original culture and at the same time become an 'all American'? What does one gain or lose in becoming completely Americanized?" Prompt 2 had higher correlation coefficients compared to Prompt 1. When looking at Prompt 2 essays I found that students reacted more to it than the first prompt. The second prompt had more guidance and helped students think more, and therefore they had more to write. The second prompt also referred to the reading prompt in the question which the first prompt did not do (although a reading prompt was provided). Though more students did Prompt 1 which was about two cultures they were familiar with, they copied from the reading prompt extensively and included little of their own ideas, but Prompt 2, the question about the gain or loss of becoming completely Americanized, seemed to get a better response from the students, and hence this prompt seemed to stimulate the students to write more. The second prompt may have resulted in a better response as students felt they had more to write about the topic, which was evident from the length of their essays. This prompt may have also tested their powers of reasoning as this probably was a topic they could identify better with as opposed to Prompt 1 where they relied on the reading prompt more than for Prompt 2.

Results for Spring 2008

Research Question 2:

Research Question 2 had the DWS teachers identifying misplaced or borderline students from the three courses of WRI 001, WRI 101 and WRI 102. The seven teachers who responded to this research question selected 53 WRI 001 students, 30 WRI 101 students and 33 WRI 102 students. However, 53 of these students came in through the system (and thus did not take the EPT in Spring 2008) and out of the 63 students remaining, 16 EPTs could not be found in the Spring 2008 EPT record, implying that they had probably taken an earlier EPT. Out of the 47 scripts only 44 could be used in this research because the other three scripts had different prompts, as they were administered at an earlier date. Among the 44 scripts, 5 students were from WRI 101 and 39 students were from WRI 001.

There appeared to be some questions in the teachers' minds as to what constitutes borderline/misplacement. According to this research, the identification of "borderline" students by teachers was when they felt these students could have benefited from placement at the next (higher or lower) level. For example, a WRI 001 student could have benefitted from being placed in WRI 101. The students themselves felt "misplaced" at the level they were in and felt they deserved to be in a class of the next level. Thus the phrase "borderline/misplaced students" had both these categories in mind. The students themselves felt they had been misplaced by the EPT, and teachers felt that some of their students could have benefitted with a higher/lower placement. There is another point that needs clarification which is that the students who felt they were misplaced were not necessarily the same students whom the teachers identified as borderline.

When I first communicated with the teachers, I told them to select the borderline students in their classes. The teachers responded with a question about who these borderline students were. Were they at the top end or the bottom end of their classes? I then asked them to identify those students who they felt could have been moved up to the next level and those they felt needed to be moved down one level – students that did not seem to fit in with the general level of the class.

Research Question 3:

The third research question dealt with the results of the holistically scored EPT in comparison with the analytically scored EPT and with Criterion^{SM} , and what these results revealed about the appropriacy of placement of these identified "borderline/potentially misplaced" students in Spring 2008.

In Spring 2008 the prompts were different from those in Fall 2007. Prompts for Spring 2008 are listed below:

> Prompt 1: Write an essay in which you explore your ideas about the cultural complexities young students face upon entering a multicultural university like the American University of Sharjah. What skills do you believe are necessary for students to have in order to function well within this culturally diverse community? Be sure to provide details and examples to support your ideas.
>
> Prompt 2: Consider the various types of technology at use now – cellphones, Ipods, Internet – and formulate your own position on how/whether they have changed people's behavior. Then, in a well-developed essay, offer your point of view on technology's impact on social interaction. Be sure to provide details and examples to support your ideas.

With regard to Research Question 3, the suggested placement for the identified borderline/potentially misplaced students, Tables 6 and 7 display the placement suggestions for Spring 2008 for Prompt 1 and Prompt 2, respectively. Since the in-class writing assignment was not available for this research, Criterion^{SM} was used as a third measure as Criterion^{SM} measures the same overall features that the holistic EPT and the analytical EPT address, albeit from a different perspective.

Results for Prompt 1

In Table 6 (Evaluation of Suggested Placements for Prompt 1 in Spring 2008) while observing the T-scale scores and comparing the holistic EPT with the analytical EPT and Criterion^{SM} scores, a difference of more than three was considered meaningful, because at AUS our grade intervals are approximately at a three-point variation on a 100-point scale for grades B and C. For example, 70–73 (4 points) is a C–, 74–76 (3 points) is a C grade, and 77–79 (3 points) is graded as a C+ while 80–83 (4 points) is a B−, 84–86 (3 points) is a B grade, and 87–89 (3 points) is graded as a B+. An A− grade is 90−94 (5 points) and an A grade is 95−100 (6 points). I compared T-scale scores between the different measures, and took any scores greater than a three-point difference between the holistic and at least one of the other two measures (analytical or Criterion^{SM}) as indicative of a possible placement differential. In case of scores showing an increase of three points or more in the analytical and Criterion^{SM} scores in comparison to the holistic EPT scores, it was considered as suitable for a higher placement. Analytical and Criterion^{SM} scores three or more points less than the holistic EPT scores were considered suitable for a lower placement. Mixed placements were suggested for those analytical and Criterion^{SM} scores less than three points different than the holistic EPT and/or where the analytical and the Criterion^{SM} scores were not in agreement with each other.

Table 6

Evaluation of Suggested Placements for Prompt 1 in Spring 2008

	Holistic EPT Score	Combined Analytical EPT Score	Criterion EPT Score	**Suggested Placement**
1	50.4307	44.4107	68.0595	–
2	54.9526	37.8897	56.5671	↓
3	36.8649	56.8597	45.0747	↑
4	54.9526	29.5904	56.5671	↓
5	54.9526	55.6740	68.0595	↑
6	45.9087	49.7459	56.5671	↑
7	54.9526	58.0453	33.5823	–
8	27.8210	53.8956	33.5823	↑
9	54.9526	56.2668	33.5823	↓
10	45.9087	55.0812	45.0747	↑
11	50.4307	49.1531	56.5671	↑
12	50.4307	42.6322	56.5671	–
13	45.9087	39.0754	45.0747	↓
14	27.8210	53.3028	45.0747	↑
15	54.9526	49.1531	56.5671	↓
16	45.9087	42.0394	56.5671	–
17	54.9526	38.4826	45.0747	↓
18	45.9087	49.1531	45.0747	↑
19	63.9965	73.4583	56.5671	–
20	63.9965	65.7518	45.0747	↓
21	63.9965	50.3387	45.0747	↓

Key to Suggested Placements for Prompt 1 in Spring 2008		Total No of 21 students
↑	Combined Analytical EPT score and/or Criterion^{SM} score higher than holistic EPT	8
↓	Combined Analytical EPT score and/or Criterion^{SM} score lower than holistic EPT	8
–	Combined Analytical EPT score and/or Criterion^{SM} score showing mixed results when compared to holistic EPT	5

While observing the results for Prompt 1, when holistic EPT scores are compared with the combined analytical EPT and Criterion^{SM} scores, eight students appear to have been placed higher than they should have been, eight students appear to have been placed lower, and five students showed a mixed result with a dichotomy in the comparison of the analytical EPT scores and the Criterion^{SM} scores when considering score differences greater than 3 points.

Another notable feature was that Criterion^{SM} EPT scores seemed to be closer to the holistic EPT scores than the third measure in twelve of the twenty-one cases in Prompt 1. In the case of nine students, however, the analytical EPT score appeared to be closer to the holistic EPT score. With regard to the suggested placements of eight of the students who were suggested to go down a level, in five of the eight cases, Criterion^{SM} EPT scores were seen to be closer to the holistic EPT scores, while only in three cases were the analytical scores closer to the holistic scores. For the eight students who required to be placed higher, five of their Criterion^{SM} scores were closer to the holistic EPT scores compared to the three cases which showed that the scores had more relation with the analytical EPT scores. For the five students who had a mixed suggested placement, in two cases Criterion^{SM} was more closely related to their holistic EPT score. However, the analytical EPT score of three of these five mixed placements were closer to their holistic EPT score.

On checking the EPT scripts for students who answered Prompt 1, I found that for the students who were placed lower in the suggested placement, their problems mainly dealt with a lack of good organization and the absence of ideas. Their paragraph structure and mechanics were not the major problem; however, their ideas seemed to be random and not connected, as seen in Student 2 and Student 4 (Sample of Student 2 and 4 attached in Appendix H). For the students whose suggested placement indicated that they should have been placed higher, it was found that handwriting, spelling, word spacing, organization of ideas, and going off the point were major areas of weakness. In Student 18's paper, word spacing was a concern (Sample of Student 18 attached in Appendix H). There seemed to be good ideas, but spelling and phraseology were key concerns. In Student 9's case the structural issues, like the length of paragraphs and the general length of the essay, were not adequate, but the ideas present in the essay were unique, simple and had a freshness about them. (Sample of Student 9's paper attached in Appendix H.) Handwriting seemed to be a distracter in Student 8's, Student 18's and Student 14's papers. (Sample of Student 8, Student 14 and Student18's paper attached in Appendix H.) Spelling appeared to be the main concern in Student number 6, 10 and 11. (Sample of Student 6, Student 10, and Student 11 attached in Appendix H.) In Student 11's paper both paragraph length and length of essay were adequate, but the phraseology, spelling and word spacing were distracters. However, Student 11 had ideas which were novel and examples which were simple enough to convey their meaning, and that is probably why this student got a higher suggested placement.

Student 18 was a case which was special, as the holistic EPT score and the Criterion^{SM} score were identical. The analytical EPT score was just a few points higher, and so this was a case which needed to be examined. Though Student 18's paper lacked in spelling, phraseology, and word spacing, the ideas of the essay seemed to be realistic and interesting. This fact of interest and realism surely would appeal to a human reader, which is perhaps why the analytical EPT score seemed to have given more weightage to ingenuity of expression, something which both Criterion^{SM} and the holistic EPT did not give attention to, and surprisingly both had a score of 45.xxxx for this student (45.9087 for the holistic EPT and 45.0747 for Criterion^{SM}). This demonstrates how Criterion^{SM} is looking at discrete points in writing and the holistic EPT is looking at the overall picture but both miss out on the content factor of which, given the nature of this research, the analytical raters involved in this study were definitely aware. This student was also another example of bad handwriting affecting the student's score and eventual placement.

In Student 8's paper handwriting was a concern. This paper was practically illegible unless very carefully read, but the writer had style, organization and correct usage. The factor which would put any rater off was the dismal handwriting. Another paper which got my attention was Student 19 who scored higher on the analytical EPT but did not do so well on Criterion^{SM}. This was one of the mixed results which showed the paper lacking in spelling, phraseology and coherence, which is what Criterion^{SM} would have picked up, but the paper did have content which Rater 1 and Rater 2 both recognized in their analytical assessment. (See Sample of Student 8, and Student 19 attached in Appendix H.)

Thus the results for Prompt 1 present a mixed picture with eight students having been placed higher, eight having been placed lower, and five being mixed in the total of twenty-one students who answered Prompt 1. For those who were placed higher and lower, the Criterion^{SM} scores were more in sync with the holistic EPT scores. For those who had mixed placements, the analytical scores were more in sync with the holistic EPT scores.

Results of Prompt 2

While observing the results for Prompt 2 in Table 7, Suggested Placements for Prompt 2 in Spring 2008, when holistic EPT scores are compared with the combined analytical EPT, seven students should have been placed higher, nine students should have been placed lower, five students had mixed placement scores and two students should remain where they were because their scores did not reflect a difference of more than three points which would have changed their suggested placement. When looking at EPT scores for those seven students who should have been placed higher, it was found that for four of these students Criterion^{SM} scores were closer than the analytical to their scores on the holistic EPT while only three of the analytical EPT scores were closer to the holistic EPT. Observing the scores of the nine students who received lower placement suggestions, it was found that holistic scores in their EPT coincided more with the Criterion^{SM} scores for six students and only for three students were their analytical EPT scores closer to their holistic EPT score. When looking at the five mixed placement suggestions, it was found that three of the analytical EPT scores were closer to the holistic EPT scores compared to only two Criterion^{SM} scores. For the two students who should have remained where they were because their placement scores did not show sufficient difference to move up or down, in one case the holistic EPT score was closer to the analytical EPT score while in the other student’s case the holistic EPT score was closer to the Criterion^{SM} score.

Table 7

Evaluation of Suggested Placements for Prompt 2 in Spring 2008

	Holistic EPT Score	Analytical Combined Score	Criterion EPT Score	**Suggested Placement**
1	40.4257	19.2745	33.7940	↓
2	50.0000	56.4685	57.8416	↑
3	50.0000	54.6393	33.7940	–
4	59.5743	60.1269	57.8416	∗
5	50.0000	61.9562	45.8178	–
6	50.0000	50.9809	45.8178	↓
7	50.0000	29.0303	45.8178	↓
8	40.4257	58.2977	57.8416	↑
9	50.0000	38.1764	45.8178	↓
10	59.5743	57.0783	57.8416	∗
11	30.8515	50.9809	57.8416	↑
12	59.5743	47.9322	45.8178	↓
13	40.4257	52.2004	57.8416	↑
14	50.0000	46.1030	57.8416	–
15	40.4257	48.5419	45.8178	↑
16	50.0000	56.4685	45.8178	–
17	50.0000	51.5906	57.8416	↑
18	50.0000	47.9322	33.7940	↓
19	59.5743	55.8588	57.8416	↓
20	50.0000	57.6880	45.8178	–
21	30.8515	42.4446	33.7940	↑
22	69.1485	52.2004	69.8654	↓
23	69.1485	54.0296	57.8416	↓

Key to Suggested Placements for Prompt 2 in Spring 2008		Total No of 23 students
↑	Combined Analytical EPT score and/or Criterion^{SM} score higher than holistic EPT	7
↓	Combined Analytical EPT score and/or Criterion^{SM} score lower than holistic EPT	9
–	Combined Analytical EPT score and/or Criterion^{SM} score showing mixed results when compared to holistic EPT	5
∗	Combined analytical and the Criterion^{SM} scores do not reflect more than a three point variation, so placement remains the same.	2

At this point an interesting pattern emerges, although the numbers are small. For students who should have been placed higher or lower, their Criterion^{SM} scores seem to be more in sync with their holistic EPT scores for both the prompts and it is the analytical scoring that picks up on the crucial differences. However, in the case of mixed placements, the analytical scores were seen to be more in sync with the holistic EPT. This reinforces the idea that there is a bit of common ground with the way the holistic EPT and Criterion^{SM} mark the papers. However, in the case of a mixed placement, the closer review factor comes into play where the analytical scoring is definitely more thorough, and with each of the components, organization, content, vocabulary, language use and mechanics being marked, the analytical scale seems like an apt additional measure to have for AUS placements.

On examining the scripts of the students' Spring 2008 EPT for Prompt 2, I found that for students whose suggested placement was lower than their holistic EPT, the general problems pertained to inadequate length of paragraphs, the essays seemed to lack structure, and the length of the essays was also a problem. All these factors were offshoots of the fact that these essays lacked development of ideas. Ideas were mentioned in the paragraphs, but they were not subsequently developed. Students who were suggested to be placed lower were the ones with no novel ideas, poor organization and weak phraseology. For students who were placed lower, namely Students 1, 6, 7, 9, 10, 12, 18, 19, and 23, the score given by Criterion^{SM} was more in sync with the holistic EPT than the analytical EPT was (sample of Students 1, 6, 7, 9, 10, 12, 18, 19 and 23 in Appendix I). In the case of Student 7 (sample in Appendix I), as can be seen, the student has no idea of a paragraph and the writing on the page has gone on from one point to the other in a haphazard manner. This student's ideas seem to be incoherent and thus provide reason enough to suggest placing the student lower than did the holistic EPT. On looking at Student 10's paper (see Appendix I), though grammar, handwriting, phraseology and mechanics seem to be in order, the use of the second person may have put the analytical rater off. Plus, the topic, instead of being analyzed, has mainly been described. This essay was full of good ideas, but the presentation of material was so sloppy that this paper was given a lower rating in the suggested placement. In the case of another paper, that of

Student 9, it was seen that though the handwriting seemed to flow, spelling errors were noticeable, and this became a distracter for the rater.

On examining the scripts for the students suggested to be placed higher, I realized that bad handwriting and incorrect spellings were their major areas of weakness. They had good ideas but bad organization. In some cases the grammar, phraseology and the vocabulary were dismal, but the ideas contained were fresh and deserved to be noticed and granted credit for. What was remarkable in three of the papers suggested to be placed higher was that Student 2, Student 8, and Student 17 (sample of Students 2, 8, and 17 attached in Appendix I) all wrote an essay that was off the point. This could have been because they misunderstood the prompt or they deliberately did not answer the prompt as it would require argumentative writing which they could not produce. However, if a student has interpreted the topic in his own way and sounds convincing, then credit should be granted to such a student. In a placement situation, test-taking pressure is not a negligible issue, and if in their nervousness students have misrepresented the question, raters must look out for factors of coherence, explaining both positions, explaining a point of view, giving a compare/contrast perspective or, better still, making the essay a descriptive piece. Raters need to consider the extent to which the student has made sense of the topic and responded in a reasonable manner. Looking at Student 11's paper (see Appendix I), though handwriting and spelling appear to be prime distracters and at times make the paper illegible, the analytical raters were able to take time to sift through these issues and arrive at the conclusion that this student does have good ideas, which could be why the suggested placement of this script landed in the higher category.

Finally, observing essays in the mixed placement section, there seemed to be numerous problems. From too many paragraphs, to random ideas and not really answering the topic, to grammar, spelling and mechanical errors, the causes for the mixed placements of these scripts seem really justifiable. Take for example, Student 22. (A sample of this essay is seen in Appendix I.) This essay has eight paragraphs, and as a result none receive their due as far as structure and coherence go. The thought process is neither unified nor showing a progression. This essay not only shows random ideas; it also is merely describing aspects of the topic rather than providing an opinion or taking a stance. Another example of the same sort was that of Student 5's paper which had ideas, but the structure, the phraseology and the subject were not adhered to (Sample of this paper in Appendix I). This could have been one of the reasons why Criterion^{SM} did not consider the essay highly since Criterion^{SM} will pick up whether or not the essay is on topic, has structure and suitable phraseology. In the case of Student 10's paper, though he had an opinion, it did not stand out as a strong one, and the essay appeared to be more descriptive than narrative.

Thus we can observe that for Prompt 2 certain factors came in to play, making these scripts to become more descriptive than argumentative. Another feature was that some students did not demonstrate an awareness of a thesis statement, supporting statements, topic statements and the general structure of the paragraph which are features of essays common in American universities. The analytical EPT definitely gave value to ideas, content, and issues of interest, which Criterion^{SM} was not specifically noting. Criterion^{SM} picks up errors of structure, organization, development and mechanics in a way that does not include the creativity and the ingenuity of a writer. It takes a human being to notice these features, which use of the analytical scale would provide.

IMPLICATIONS

The first research question of this research was regarding *a priori* and *a posteriori* evidence about the validity and the reliability of the proposed analytical scale for use with the EPT in the Fall 2007 semester. The second question sought to identify the "borderline/potentially misplaced" students, if any, in the WRI courses for the Spring 2008 semester, and the third question investigated the appropriacy of placement of the identified "borderline/potentially misplaced" students, by comparing their analytical, holistic and Criterion^{SM} EPT scores.

Implications for Placement Testing

A notable feature of this research was the very high inter-rater reliability of 0.900 on an average for all the 109 scripts analyzed, both in Fall 2007 and Spring 2008. In fact, Cronbach's alpha for Prompt 2 in Fall 2007 was almost perfect at 0.990, while the lowest inter-rater reliability was 0.970 for Prompt 1 in Spring 2008.

There was a difference, however, between the Prompt 1 and Prompt 2 in Fall 2007, resulting in lower correlations for Prompt 1, which raises a very important question about the suitability of the prompt both for students and raters. The wording of the prompt and its relevance to the reading text were two issues which may have significantly influenced students' performance. For example, Prompt 1 in Fall 2007 dealt with considering two cultures that students were familiar with. The prompt asked them to write about the differences in communication styles in these two cultures. The reading prompt was not referred to in this question, and students had to really think to answer. Prompt 2, on the other hand, had reference to the reading prompt. The character in the reading prompt experienced a clash between two cultures, and the question the students were asked about their feelings about becoming "Americanized." This topic seemed to appeal to students more as they could probably identify with it more, and plus, there was more direction as to what to write on. Thus selecting the prompt and paying special attention to the essay question asked about a reading text are areas which need attention in future EPTs. This is an area which administrators need to consider when prompts are being selected for the placement tests.

The scatterplots of Fall 2007 all showed linear relationships between all scorings as seen in Appendix E and Appendix F. This was to be expected because the scripts used in the research were a representative sampling of WRI 001, WRI 101 and WRI 102 and because the holistic, analytical and Criterion^{SM} scorings tapped into the same construct, as demonstrated in the moderately strong overlapping variance.

While considering the Pearson's correlation coefficient, the results of Fall 2007 show the correlation was the strongest between the holistic and the analytical scoring of Prompt 2 (r = 0.901). This indeed is very meaningful because it shows that, provided the prompt is carefully and suitably chosen, the analytical scoring pattern does correlate with and also complements the holistic EPT scores, demonstrating that it can be used as an additional source of evidence with regard to the "borderline/potentially misplaced" students. Prompt 1, however, showed a moderately strong correlation of 0.850 between the holistic and the analytical scores, which implies that the prompt plays a crucial role in the correlation between the holistic and analytical EPT scores. With regard to overlapping variance, in Fall 2007 Prompt 2 showed a greater extent of overlap between the holistic and the analytical scoring at 90.1%, while the Criterion^{SM} and the analytical had an overlapping variance of 65.0% between them. Why was the overlap between the holistic and the analytical EPT higher compared to Criterion^{SM} and analytical? Both the holistic and the analytical EPT had human raters, and handwriting was a concern for both these EPTs while Criterion^{SM} rated typed essays which eliminated handwriting as a factor. Prompt 1 did not record such a high overlapping variance, scoring a 72.9% between holistic and analytical and an even lower variance of 44.0% between Criterion^{SM} and analytical scoring, very likely indicating that the human rater evaluation was affected by the prompt more than the Criterion^{SM} evaluation.

This research also highlighted the issue of variations between raters, which can adversely affect placement results. Raters need to be extensively trained to carry out effective placements. In this research Cronbach's Alpha between Rater 1 and Rater 2 was quite high. Norming sessions before correcting EPT scripts should be a must to ensure that all the raters are measuring the same constructs in a similar manner. This is well reflected in the results in Fall 2007 through Pearson's moment correlation coefficient, correlating the individual analytical rater scores with Criterion^{SM}. Correlating Prompt 1's

Criterion^{SM} and analytical scores of Raters 1 and 2, Rater 1-Criterion^{SM} was at 0.614, whereas Rater 2-Criterion^{SM} was at 0.695. Even when observing the holistic and analytical correlation, Rater 1- holistic EPT was at 0.809 and Rater 2-holistic EPT was at 0.875. For Prompt 2, the correlation between Criterion^{SM} and analytical scoring between Raters 1 and 2 showed more disparity than between holistic and analytical. One possible explanation for this could be that Criterion^{SM} looks only at style, organization and mechanics but human raters also look at content and presentation. For Criterion^{SM} and analytical, the correlation for Rater 1-Criterion^{SM} was 0.790 and for Rater 2- Criterion ^{SM} was 0.812, whereas for holistic and analytical, the correlation for Rater 1-holistic EPT was 0.934 and Rater 2-holistic EPT was 0.953.

This research only measured the correlation of the holistic EPT with the analytical EPT and did not seek to test equivalency at any point. The analytical EPT did have a high correlation with the holistic EPT for Prompt 2 in Fall 2007 with a representative sampling of WRI 001, 101 and 102 students, which does indicate that it measured the EPT scripts similarly to faculty evaluation and thus it may be appropriate to use as an additional measure in the case of "borderline/potentially misplaced" students.

Observing the results for Prompt 1 and Prompt 2 in Spring 2008, in the case of the higher and lower placements, Criterion^{SM} was more in sync with the holistic EPT, and it was the analytical scoring that picked up on distinguishing features affecting placement. The surprising feature was that for the mixed cases, the analytical EPT scores were more in sync with the holistic EPT scores, and Criterion^{SM} picked up on features of student writing that the other two measures did not. So for most of the students whose suggested placements were higher or lower, it was the analytical scale which identified the misplacement. Ultimately, Criterion^{SM} measured form and structure, much like the holistic EPT rater who would rate the papers quickly, looking at an overall picture of form and structure rather than each individual feature of writing and would miss good content obscured by bad handwriting, etc.

In the case of mixed placements, the human element and handwriting seemed to play a huge role, and out of the ten cases of mixed placements, in six of these scripts, the analytical score is more closely related to the holistic EPT score rather than Criterion^{SM} which was not affected by handwriting and less so by spelling. In the case of sure suggested replacing, be it high or low, the analytical EPT was able to pick up certain aspects of the students' writing which the holistic EPT was not able to see. The rubric for the analytical scale has content, organization, vocabulary, language use and mechanics. It is this breakdown of features which was able to pick up certain aspects of student writing which the holistic EPT missed.

This pattern can be noticed with EPT Criterion^{SM} scores, too. In the cases of suggested replacement, being high or low, where the analytical EPT score was not in agreement with the holistic EPT, Criterion^{SM} EPT scores were seen to be more in agreement with the holistic score. For example, in Spring 2008 for Prompt 1 and 2, out of the total fifteen cases, in ten of the cases Criterion^{SM} is more closely related to the holistic EPT. Criterion^{SM} was picking up different aspects of writing than those being observed by the analytical EPT. It was the analytical evaluation which identified the misplacement. This observation implies that the analytical scoring of the EPT can be used as an additional measure to verify student placement. Criterion^{SM} uses a holistic score similar to holistic EPT scoring, and Criterion^{SM} focuses on form, structure, etc., more than content which would be picked up by human raters.

Spelling and handwriting emerged as other major factors affecting student placements. While closely examining the EPT scripts of the students, I found that in papers where there was a marked difference in the grading of the holistic EPT and the analytical EPT, handwriting and spelling played a crucial role. Raters in DWS were probably looking at presentation, and spelling and handwriting appeared key in their evaluating of student papers. The bad handwriting feature was eliminated completely while examining the same script on Criterion^{SM}, as all essays were typed. This handwriting factor could have been one of the reasons for student misplacement. If students were to use computers in the labs at AUS, this problem of bad handwriting could be avoided as it is a major problem with most students coming to AUS. Spelling could be still monitored if the spell check feature was removed, or alternatively, it could be available because there is still a choice of words which the spell check gives and students would still have to choose the correct word from the list. If hand written papers are going to be the norm, then students must be told in the instructions given to them at the beginning of the placement test that they should pay special attention to their handwriting and spelling as it would be a major consideration while their papers are being scored.

The dichotomy in the time difference between the higher educational requirements on Criterion^{SM} and the AUS EPT was another implication which needed to be considered. The higher education time requirement used by ETS in the Criterion^{SM} evaluation is thirty minutes as against that of the hour-long AUS EPT. Criterion^{SM} topics also do not include a reading prompt, indicating a difference in prompt attributes and test task characteristics even though the response attributes are the same (which are what is being measured).

The other important issue to be considered with respect to placements is that all future EPTs will be conducted by AUS personnel in the Testing Center, and the DWS faculty will not administer the exam. This means that there will be a variety of prompts and the scripts will be marked by different raters at different times during the semester, raising questions of the reliability and validity of accurate placement, as was discovered in the case of Prompt 1 and 2 in Fall 2007. The Department of Writing Studies could develop a test bank of tried and tested topics which could be used for future EPTs. There should be rater training and norming sessions before the EPT each semester to ensure that all the teachers involved are on the same page with regard to marking the EPT. The EPTs should be scheduled consistently, and the Department of Writing Studies needs to be informed well in advance so that the prompts can be worked out, tried and tested before they are given to a group of new admissions. There needs to be a team involved in selecting the prompts for the EPTs, and this team should be preferably headed by a testing expert. The selection of the prompts need special attention as was revealed through this study, which can only be done by having a test pool and establishing test specifications.

Limitations of This Study

There are several limitations to this study. During the Fall 2007 semester I only looked at 65 EPT scripts to evaluate the use of the analytical scale with the EPT. When the total number of students that take the EPT ranges from 400 to 800, sampling only 109 scripts (65 scripts in Fall 2007 and 44 scripts in Spring 2008) seems to be a small basis for analysis to validate the analytical scale for use in the case of "borderline/potentially misplaced" students. In particular, the small sample size for Prompt 2 in Fall 2007 could be problematic with Pearson which is more effective with large sample sizes. Ideally, all EPTs should have been used to correlate the analytical scaled EPTs with the holistic scaled EPT and Criterion^{SM}, but doing so was not practical. Also, there were logistical difficulties with double marking 65 scripts in the Fall 2007 semester and the subsequent typing of all these handwritten EPT essays to put them through Criterion^{SM}. However, even though only 65 scripts were scored in Fall 2007, they represented all three classes of WRI 001, WRI 101 and WRI 102. The issue of different prompts in Fall 2007 and Spring 2008 was another limitation.

Also, I was only evaluating student placement as done at the beginning of the Spring 2008 semester. Because of time constraints, it was not possible to track the participating students' progress in their classes throughout the semester to investigate further the predictive validity of the analytically scored EPT.

Final Thought

The Department of Writing Studies faculty are already burdened with a large number of writing courses, and this research does not wish to make life any more difficult than it already is. To the contrary, I want to suggest an additional measure so that the teachers have added resources at hand when students complain about being misplaced or when a teacher feels a student is a borderline case in his/her class. If the analytical scale can be used as an additional measure in the case of "borderline/potentially misplaced students," this would lighten the burden of teachers who have no way to address misplacement after the EPT and their in-class assignments are over. Criterion^{SM}, too, could prove useful since it is less affected by spelling and handwriting problems. However, simply having students' type their essays would address these problems in part, at least without incurring the subscription costs of Criterion^{SM}. These alternatives can thus prove effective in addressing these " borderline/potentially misplaced " cases.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Attali, Y. (2004). *Exploring the feedback and revision features of Criterion.* Paper presented at the National Council on Measurement in Education (NCME), April 12 to 16, 2004, in San Diego, CA. Retrieved October 3, 2007, from www.ets.org/

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *The Journal of Technology, Learning and Assessment, 4*(3), 1-31. Retrieved July 24, 2008, from http://www.jtla.org

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671-704. Retrieved July 20, 2008, from the JSTOR database.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42. Retrieved July 20, 2008, from the Ebscohost database.

Bachman, L. F. (2005). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

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

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions and directions.* Boston: Heinle & Heinle Publishers.

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment, 6*(1), 1-47. Retrieved July 24, 2008, from http://www.jtla.org

Bowen, J. D., Madsen, H., & Hilferty, A. (1985). *TESOL techniques and procedures.* Cambridge, MA: Newbury House Publishers.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices.* New York: Longman.

Brown, J. D. (1997). Reliability of surveys. *Shiken: JALT testing and evaluation SIG newsletter, 1*(2), 17-19. Retrieved July 24, 2008, from http://www.jalt.org/test/editors.htm

Brown, J. D. (2001). *Using surveys in language programs.* Cambridge: Cambridge University Press.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653-675. Retrieved July 16, 2008, from the JSTOR database.

Chalhoub-Deville, M. (1999). *Studies in language testing: Issues in computer-adaptive testing of reading proficiency.* Cambridge: Cambridge University Press.

Clark, I. L. (2003). *Concepts in composition: Theory and practice in the teaching of writing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, A. D. (1994). *Assessing language ability in the classroom* (2nd ed.). Boston: Heinle and Heinle.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 compositions ratings really mean? *TESOL Quarterly*, *29*(4), 762-765. Retrieved July 24, 2008, from the JSTOR database.

Coombe, C. A., & Hubley, N. (2003). *Assessment practices.* Alexandria, VA: Teachers of English to Speakers of Other Languages, Inc.

Crusan, D. (2002). An assessment of ESL writing placement assessment. *Assessing Writing, 8*(1), 17-30. Retrieved July 24, 2008, from the Science Direct database.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Studies in language testing: Dictionary of language testing.* Cambridge: Cambridge University Press.

Dodigovic, M. (2005). *Artificial intelligence in second language learning: Raising error awareness.* Clevedon: Multilingual Matters Ltd.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-187. Retrieved July 20, 2008, from http://ltj.sagepub.com/cgi/content/abstract/25/2/155

Educational Testing Service. (2007). Retrieved September 20, 2007, from http://www.ets.org

Electronic references: Criterion scoring guide. (2007). *Criterion.ets.org* . Retrieved February 22, 2009, from http://www.ets.org/Media/Products/Criterion/topics/topics.htm

Freeman, D. (Ed.). (1998). *Doing teacher research: From inquiry to understanding.* Toronto: Heinle & Heinle Publishers.

Garcia-Mayo, M. P. (1996). The reliability of the holistic method when grading language essays. *Cuadernos de Filologia Inglesa, 5*(1), 51-62. Retrieved July 20, 2008, from http://www.dialnet.unirioja.es/servlet/fichero_articulo?codigo=1325557&orden=0

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies 1992-2002. *JTLA The Journal of Technology, Learning and Assessment, 2*(1), 1-52. Retrieved July 24, 2008, from http://www.jtla.org

Green, A. B., & Weir, C. J. (2004). Can placement test inform instructional decisions? *Language Testing, 21*(4), 467-494. Retrieved July 20, 2008, from http://ltj.sagepub.com/cgi/content/abstract/21/4/467

Hamp-Lyons, L. (1990). *Assessing second language writing in academic contexts.* Connecticut: Ablex Publishing.

Hamp-Lyons, L. (1995a). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762. Retrieved July 16, 2008, from the JSTOR database.

Hamp-Lyons, L. (1995b). Uncovering possibilities for a constructivist paradigm for writing assessment. *College Composition and Communication, 46*(3), 446-455. Retrieved July 16, 2008, from the JSTOR database.

Harmer, J. (2001). *The practice of English language teaching* (3rd ed.). Essex, England: Pearson Education Ltd.

Harris, D. P. (1969). *Testing English as a second language*. New York: Mc Graw-Hill.

Haswell, R., & Wyche-Smith, S. (1994). Adventuring into writing assessment. *College Composition and Communication, 45*(2), 220-236. Retrieved July 16, 2008, from the JSTOR database.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Huot, B. (1990a). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237-263. Retrieved July 20, 2008, from the JSTOR database.

Huot, B. (1990b). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201-206. Retrieved July 24, 2008, from the JSTOR database.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfield, V.F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach.* Rowley, Mass: Newbury House.

Kroll, B. (1990). *Second language writing: Research insights for the classroom.* Cambridge: Cambridge University Press.

Larson-Freeman, D. (1985). Considerations in research design. In M. Celce-Murcia (Ed.), (1985). *Beyond basics: Issues and research in TESOL* (pp. 125-133). Massachusetts: Newbury House Publishers, Inc.

Lloyd, D., Davidson, P., & Coombe, C. (Eds.) (2005). *Fundamentals of language assessment: A practical guide for teachers in the Gulf.* Dubai: TESOL Arabia.

Madsen, H. S. (1983). *Techniques in testing.* New York: Oxford University Press.

McNamara, T. (2000). *Oxford introductions to language study: Language testing.* Oxford: Oxford University Press.

Morris, T. (n.d.) . Classroom practice report: Holistic scoring/rubrics. Retrieved July 24, 2008, from http://www.case.edu/artisci/engl/emmons/writing/asgments/Morris-holistic.pdf

Moser, A. (2008). Multiple intelligences and writing traits: *Proceedings of the Daejeon-Chuncheong KOTESOL Symposium. September 27, 2008*. Retrieved July 20, 2008, from http://www.kotesol.org/?q=node/257

Nakamura, Y. (2004, May 22). The interface between inter language, pragmatics and assessments. *Proceedings of the 3rd Annual JALT Pan-SIG conference,* pp. 45-52. Retrieved July 20, 2008, from http://www.jalt.org/pansig/2004/index.html

Norris, J. M., Brown, J.D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* Manoa: University of Hawai'i.

O'Malley, J. M., & Pierce, L. V. (1996). *Authentic assessment for English language learners*. New York: Addison Wesley.

Purnell, R. B. (1982). A survey of testing of writing proficiency in college: A progress report. *College Composition and Communication, 33*(4), 407-410. Retrieved July 20, 2008, from the JSTOR database.

Reid, J. M. (1993). *Teaching ESL writing.* Englewood Cliffs, CA: Prentice Hall Regents.

Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning, 46*(1), 137-174.

Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluations. *The English Journal*, *74*(5), 49-55. Retrieved July 20, 2008, from the JSTOR database.

White, E. M. (1984). Holisticism. *College Composition and Communication*, *35*(4), 400-409. Retrieved July 20, 2008, from the JSTOR database.

Appendix A

(Sample English Placement Test)

American University of Sharjah

Department of Writing Studies

ENGLISH PLACEMENT TEST

Fall 2007

Name: ____

ID Number: ____

TOEFL Score

Computer Based: ____

Paper Based: ____

Check One:

_____New Student

_____IEP Student

_____Returning / Former IEP Student

Check (School / College)

□ College of Arts and Sciences

□ School Of Engineering

□ Undeclared Major

□ School Of Architecture and Design

□ School Of Business and Management

<u>Do not write below this line</u>____

R1	R2	R3	Total

Length Requirement: Write an essay that consists of several well-developed paragraphs.

Structural Requirements: Your essay needs to include an introduction, a thesis statement, an appropriate number of support paragraphs, and a conclusion. Content, coherence, clarity of expression, grammar, punctuation and readability will be taken into consideration when assessing your essay. Be sure to leave yourself ample time to proofread your work.

Formatting Requirements: Please write on every other line to allow yourself space for revisions and to enhance the readability of your essay.

Duration of Exam: 1 hour

"The Struggle to Be an All-American Girl" by Elizabeth Wong

It's still there, the Chinese school on Yale Street where my brother and I used to go. Despite the new coat of paint and the high wire fence, the school I Knew ten years ago remains remarkably, socially the same.

Every day at 5 p.m., instead of playing with our fourth- and fifth-grade friends or sneaking out to the empty lot to hunt ghosts and animal bones, my brother and I had to go to the Chinese School. No amount of kicking, screaming, or pleading could dissuade my mother, who was solidly determined to have us learn the language of our heritage.

Forcibly, she walked us the seven long, hilly blocks from our home to school, depositing our defiant tearful faces before the stern principal. My only memory of him is that he swayed on his heels like a palm tree, and he always clasped his impatient twitching hands behind his back. I recognized him as a repressed maniacal child killer, and knew if we ever saw his hands we'd be in big trouble.

We all sat in little chairs in an empty auditorium. The room smelled like dirty closet. I hated that smell. I favored crisp new scents. Like the soft French perfume that my American teacher wore in public school.

There was a stage far to the right, flanked by an American flag and the flag of the Nationalist Republic of China, which was also red, white and blue but not as pretty.

Although the emphasis at the school was mainly language-speaking, reading and writing- the lesson always began with an exercise in politeness. With the entrance of the teacher, the best student would tap a bell and everyone would get up, kowtow, and chant, "Sign san ho," the phonetic for "How are you, teacher?"

Being ten years old, I had better things to learn than ideographs copied painstakingly in lines that ran right to left from the tip of a moc but, a real ink pen that had to be held in an awkward way if blotches were to be avoided. After all, I could do the multiplication tables, name the satellites of Mars, and write reports on Little Women and Black Beauty. Nancy Drew, my favorite book heroine, never spoke Chinese.

The language was source of embarrassment. More times than not, I had tried to disassociate myself from the nagging loud voice that followed me whenever I wandered in the nearby American supermarket outside Chinatown. The voice belonged to my grandmother, a fragile woman in her seventies who could outshout the best of the street vendors. Her humor was raunchy, her Chinese rhythmless, patternless. It was quick, it was loud, it was unbeautiful. It was not like the quiet, lilting romance of the French or the gentle refinement of the American South. Chinese sounded pedestrian public.

In Chinatown, the comings and goings of hundreds of Chinese on their daily tasks sounded chaotic and frenzied. I did not want to be thought of as mad, as talking gibberish. When I spoke English, people nodded at me, smiled sweetly, said encouraging words. Even the people in my culture would cluck and say that I'd do well in life. "My, doesn't she move her lips fast," they would say, meaning that I'd be able to keep up with the world outside Chinatown.

My brother was even more fanatical than I about speaking English. He was especially hard on my mother, criticizing her, often cruelly, for her pidgin' speech—smatterings of Chinese scattered like chop suey in her conversation. "It's not 'What it is,' Mom," he'd say in exasperation. "It's 'What is it, what is it, what is it!'" Sometimes Mom might leave out an occasional "the" or "a," or perhaps a verb of being. He would stop her in mid-sentence: "Say it again, Mom. Say it right." When he tripped over his own tongue,' he'd blame it on stumbled in speaking her: "See, Mom, it's all your fault. You set a bad example."

What infuriated my mother most was when my brother cornered her on her consonants, especially "r." My father had played a cruel joke on Mom by assigning her an American name that her tongue wouldn't allow her to say. No matter how hard she tried, "Ruth" always ended up "Luth" or "Roof."

After two years of writing with a moc but and reciting words with multiples of meanings, I finally was granted a cultural divorce. I was permitted to stop Chinese school.

I thought of myself as multicultural. I preferred tacos to egg rolls; I enjoyed Cinco de Mayo more than Chinese New Year.

At last, I was one of you; I wasn't one of them.

Sadly, I still am.

Wong describes the clash of two cultures and the conflicts that can occur from it. Do you think it is Possible for someone to maintain connections with his or her original culture and at the same time become an "all-American"? What does one gain or lose in becoming completely Americanized?

Appendix B
(Sample of In-class Writing Assignment)

Write a paragraph about your development as a reader/writer. When did you start reading/writing? Who motivated you to read/write? Did you face any problems while reading/writing? What were some of the problems you faced initially? What problems do you still battle with now? What kind of reading/writing would you say is your favorite?

Appendix C
(Scoring the English Placement Test)

English Placement Test Grading Rubric

Possible marks on the EPT range from 1-6. As **all exams are double-marked, the final score for an exam will range between 2-12.**

001 Placement

Readers record a score of 1 or 2 to an exam which *lacks*

- clear sentence level grammar and/or mechanical competence (although spelling should not be the only factor preventing a student from placement in Com 101).
- paragraph unity and development.
- An understanding and/or an appropriate response to the prompt.

101 Placement

Readers record a score of 3 or 4 to an essay which *possesses*

- clear sentence level grammar and/or mechanical competence.
- paragraph unity and development *and a* central governing idea but may lack a clear method of organization.
- ability to understand and/or appropriately respond to the prompt.

102 Placement

Readers record a score of 5 or 6 to an essay which demonstrates an exceptional ability to compose an essay which *possesses*

- basic essay organization (introduction, body paragraphs, conclusion)
- a clear thesis..
- thesis support.
- responds meaningfully to the prompt and/or shows early signs of written reasoning or critical analysis.
- sophisticated vocabulary, appropriate word form.

Individual Score	Combined Score	Course Placement
1-2	0-5	001
3-4	6-9	101
5-6	10+	102

English Placement Test Grading Information

<u>Format</u>

• The test consists of a short reading and a prompt, a copy of which is included in this packet.

• Students will use photocopied packets that include a title page, a test sheet including instructions, reading, and prompt, and four lined-pages for writing their essay. Extra paper will be made available during the exam for those students who need more than the four pages provided.

• The title page will be used by readers to record marks and course placement.

• Time allocated for the test: 1 hour

<u>Scoring</u>

• Each individual reader will grade the test using a 6 point scale:

o A score of I or 2 indicates the reader assesses the student's writing abilities for placement in 001

1-2 -+ placement in 001

o A score **of 3 or 4 indicates the reader assesses the student's writing abilities for placement in 101**

3-4 -+ placement in 101

o A score of **5 or 6 indicates the reader assesses the student's writing abilities for placement in 102**

5-6 --* placement in 102

• All tests will be double-marked using the following system: o RI will read the exam and indicate a 1-6 score.

o Without looking at the previous reader's score, R2 will also indicate a 1- 6 score after reading the exam.

o R2 will then compare the two independent scores. If the scores are the same or adjacent numbers, R2 adds the scores and enters the overall grade.

■ test-takers earning a combined score up to and including a 5; will place into 001 • 0-5 -+ placement in 001

■ test-takers earning a combined score of 6 -9 will place into 101 • 6-9 -placement in 101

■ test-takers earning a combined score of 10 or better will place into 102 • 10 + -placement in 102

o In the case where the R1 and R2 score are not the same or adjacent, the test will go to a third reader. The non-similar score will be disregarded.

o When the final score is achieved either through two or three readings the final reader will total the numbers, record the total score in the box labeled "total", and indicate the course number (001, 101, 102) into which the student will be placed in the top right hand corner of the title page in the box labeled "course placement."

Administration

• Instructors are encouraged to work with a partner to exchange exam packets.

• All exams must be double-marked, separated into stacks according to course placement, and returned to Hadeel's office no later than Monday at 5:00 pm

Thank you all for your help!

Appendix D
(Analytical Scale)

ESL COMPOSITION PROFILE		
SCORE	LEVEL	CRITERIA
CONTENT	30-27	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic
	26-22	GOOD TO AVERAGE: some knowledge of subject • adequate range limited development of thesis • mostly relevant to topic, but lacks detail
	21-17	FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic
	16-13	VERY POOR: does not show knowledge of subject • non-substantive not pertinent • OR not enough to evaluate
ORGANIZATION	20-18	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/ 2 supported • succinct • well-organized • logical sequencing • cohesive
	17-14	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas standout• limited support • logical but incomplete sequencing
	13-10	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development
	9-7	VERY POOR: does not communicate • no organization • OR not enough to evaluate
VOCABULARY	20-18	EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register
	17-14	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage but meaning nor obscured
	13-10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • meaning confused or obscured
	9-7	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate
LANGUAGE USE	25-22	EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pro-nouns, prepositions
	21-18	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured
	17-11	FAIR TO POOR: major problems in simple/complex constructions frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • meaning confused or obscured
	10-5	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate
MECHANICS	5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing
	4	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured
	3	FAIR TO POOR: major problems in simple/complex constructions frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • meaning confused or obscured
	2	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate
TOTAL SCORE	READER	COMMENTS

Appendix E
(Criterion^{SM} Descriptors)

ETS Criterion. Online Writing Evaluation

Scoring GUIDE

Score = 6

Excellent

- Develops ideas well and uses many specific, relevant details throughout the essay.
- Is well organized with clear transitions; maintains focus.
- Sustains varied sentence structure.
- Exhibits many specific word choices.
- Contains little or no errors in grammar and conventions; errors do not interfere with understanding.

Persuasive mode

- Clearly states the position and effectively persuades the reader of validity of argument.

Score = 5

Skillful

- Develops ideas with some specific, relevant details.
- Is clearly organized; information is presented in an orderly way, but essay may lack transitions.
- Exhibits some variety in sentence structure.
- Displays some specific word choices.
- May contain some errors in grammar and conventions; errors do not interfere with understanding.

Persuasive mode

- Clearly states the position and persuades the reader.

Score = 4

Sufficient

- Provides clear ideas, but sparsely developed; may have few details.
- Provides a clear sequence of information; provides pieces of information that are generally related to each other.

- Generally has simple sentences; may exhibit uneven control over sentence structure.
- Consists mainly of simple word choices, but may contain some specific word choices.
- Contains errors in grammar and conventions that generally do not interfere with understanding.

Persuasive mode

- States a position and adequately attempts to persuade the reader.

Score = 3

Uneven

- Provides limited or incomplete information; may be list-like or have the quality of an outline.
- Is disorganized or provides a disjointed sequence of information.
- Exhibits uneven control over sentence structure.
- May have some inaccurate word choices.
- Contains errors in grammar and conventions that sometimes interfere with understanding.

Persuasive mode

- While a position is stated, either it is unclear or undeveloped.

Score = 2

Insufficient

- Provides little information and makes little attempt at development.
- Is very disorganized or too brief to detect organization.
- Exhibits little control over sentence structure.
- Contains inaccurate word choices in much of the essay.
- Is characterized by misspellings, missing words, and incorrect word order; errors in grammar and conventions are severe enough to make understanding very difficult in much of the essay.

Persuasive mode

- Either a position is not clearly given or little attempt is made at persuasion.

Score = 1

Unsatisfactory

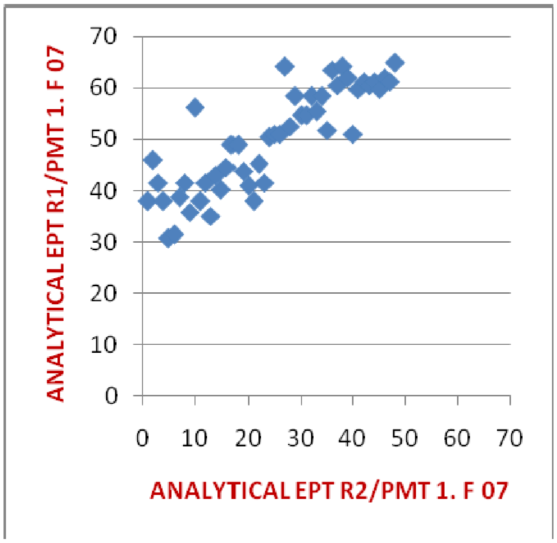
- Attempts a response, but may only paraphrase the prompt or be extremely brief.
- Exhibits no control over organization.
- Exhibits no control over sentence structure.
- Contains inaccurate word choices throughout most of the essay.
- Is characterized by misspellings, missing words, and incorrect word order; errors in grammar and conventions severely impede understanding throughout the essay.

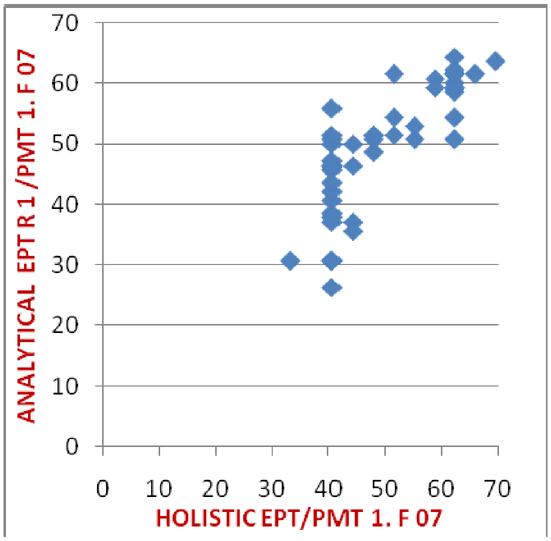
Persuasive mode

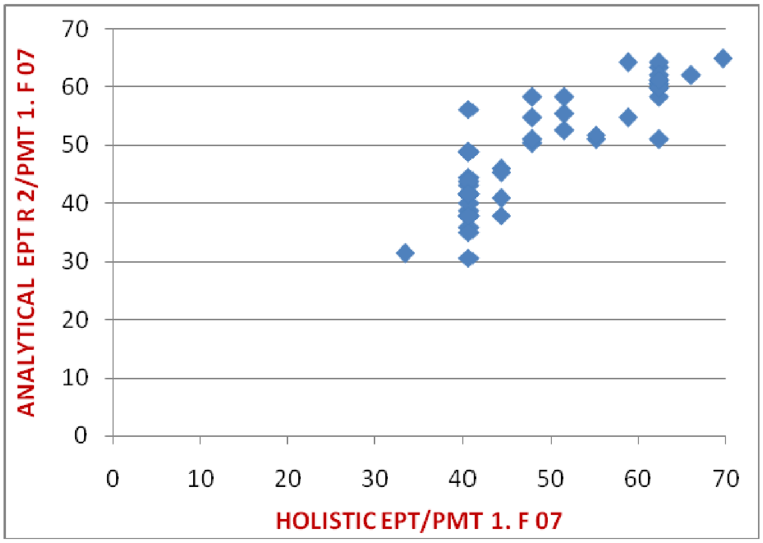
- Little effort is made to persuade, either because there is no position taken or no support is given.

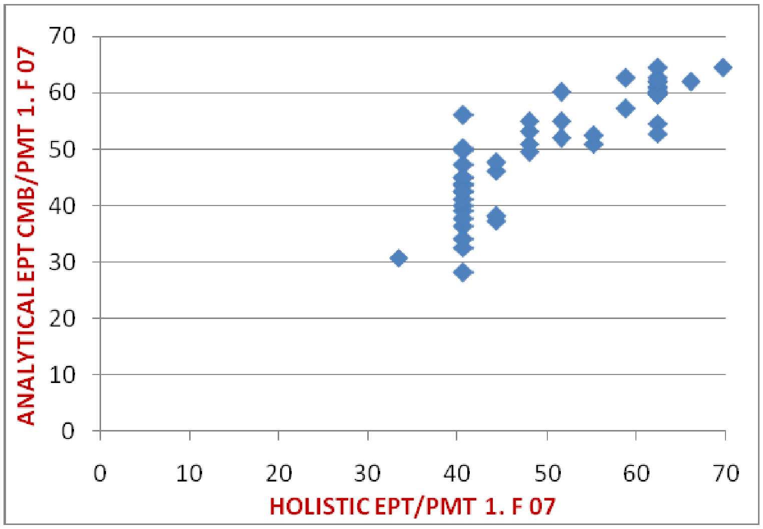
“Electronic References,” (2007) Retrieved February 22, 2009, from http://www.ets.org/Media/Products/Criterion/topics/topics.htm

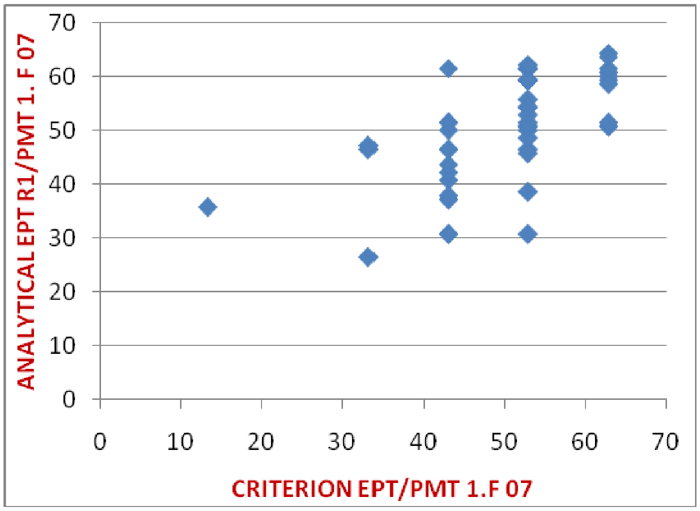
Appendix F
(Scatterplots for Prompt 1 in Fall 2007)

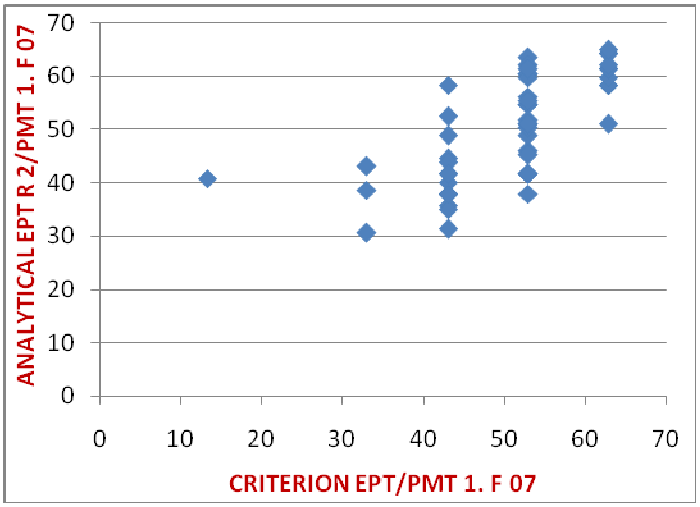

ANALYTICAL EPT R1/PMT 1. F 07
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
ANALYTICAL EPT R2/PMT 1. F 07

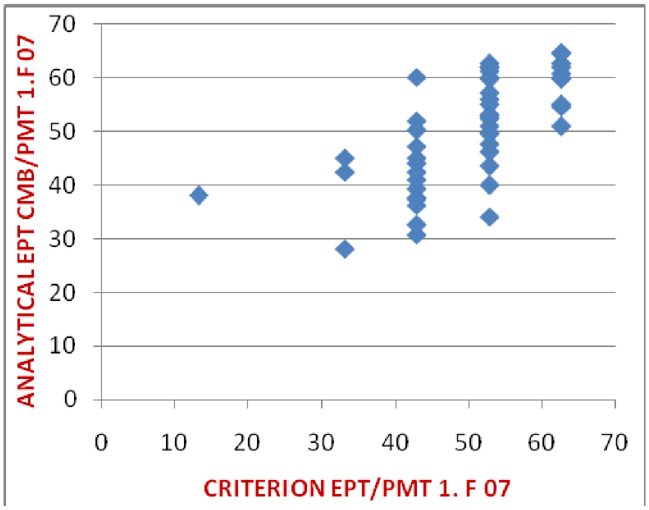

ANALYTICAL EPT R 1 /PMT 1. F 07
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
HOLISTIC EPT/PMT 1. F 07


ANALYTICAL EPT R 2/PMT 1. F 07
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
HOLISTIC EPT/PMT 1. F 07

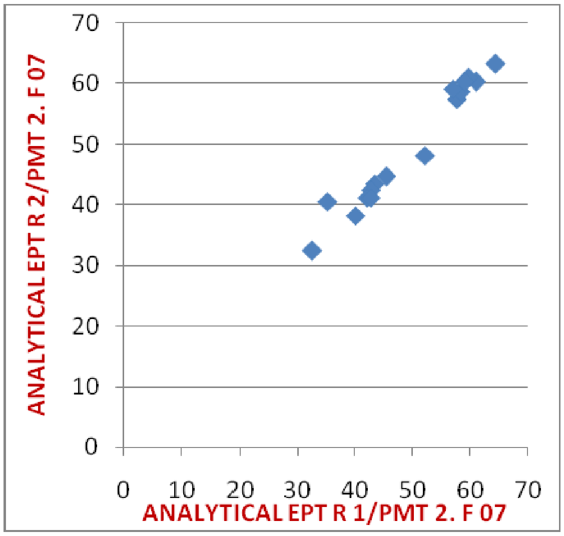

ANALYTICAL EPT CMB/PMT 1. F 07
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
HOLISTIC EPT/PMT 1. F 07

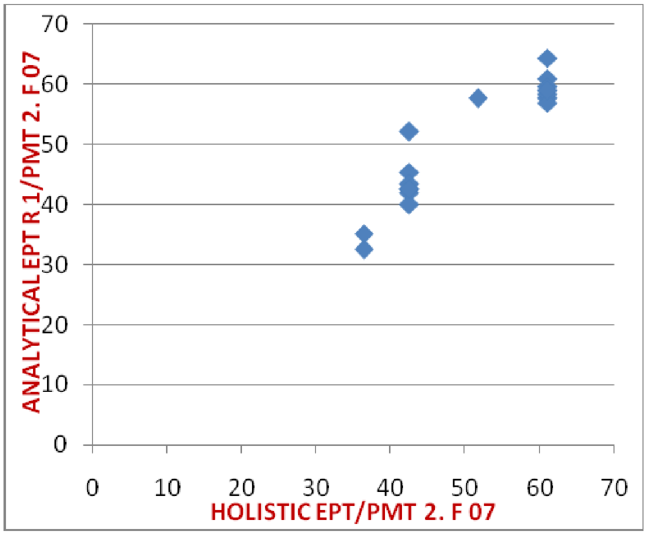

ANALYTICAL EPT R1/PMT 1. F 07
0
10
20
30
40
50
60
70
CRITERION EPT/PMT 1.F 07

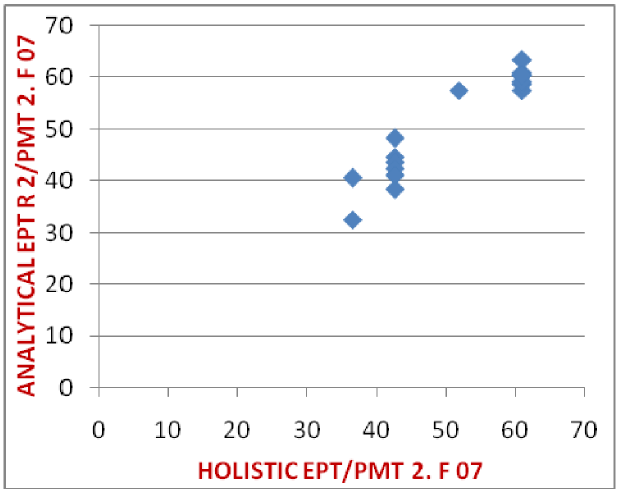

ANALYTICAL EPT R 2/PMT 1. F 07
0
10
20
30
40
50
60
70
CRITERION EPT/PMT 1. F 07

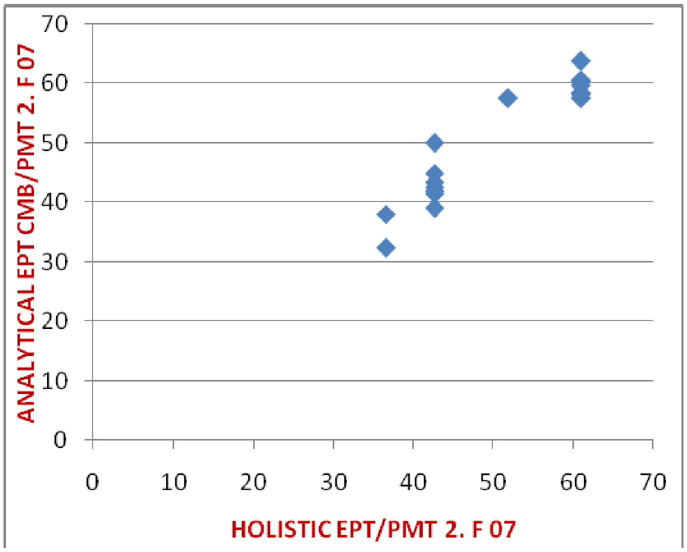

ANALYTICAL EPT CMB/PMT 1.F 07
0
10
20
30
40
50
60
70
CRITERION EPT/PMT 1. F 07

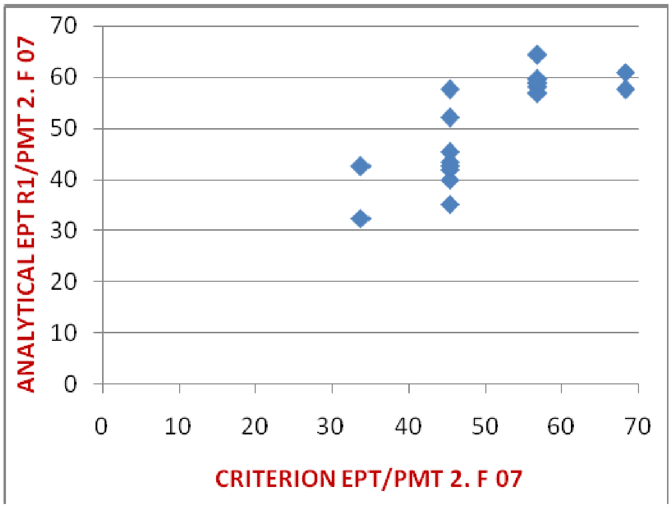
Appendix G
(Scatterplots for Prompt 2 in Fall 2007)

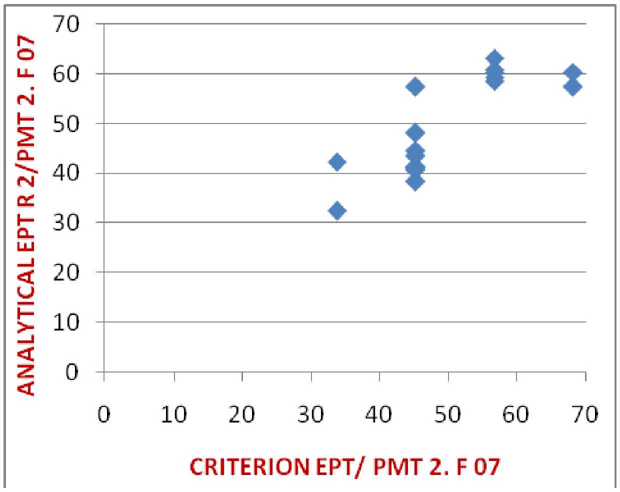

ANALYTICAL EPT R 2/PMT 2. F 07
70
60
50
40
30
20
10
0
0 10 20 30 40 50 60 70
ANALYTICAL EPT R 1/PMT 2. F 07

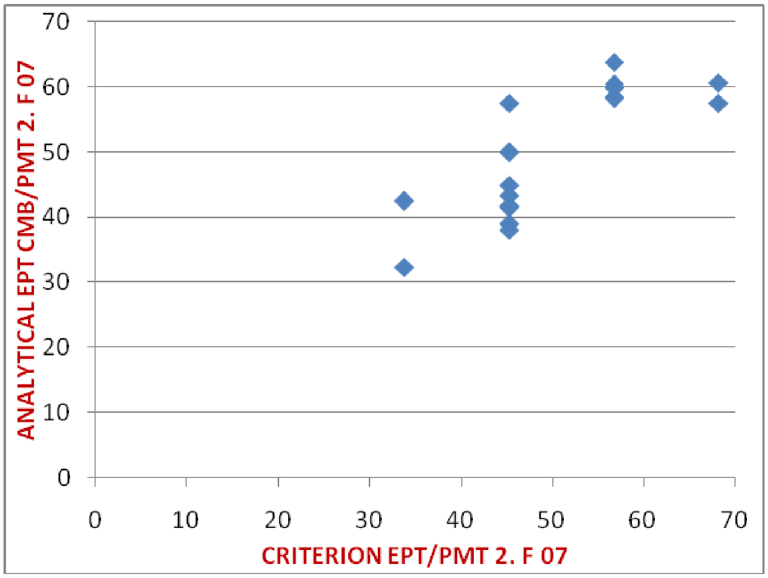

ANALYTICAL EPT R 1/PMT 2. F 07
70
60
50
40
30
20
10
0
0 10 20 30 40 50 60 70
HOLISTIC EPT/PMT 2. F 07


ANALYTICAL EPT R 2/PMT 2. F 07
70
60
50
40
30
20
10
0
0 10 20 30 40 50 60 70
HOLISTIC EPT/PMT 2. F 07

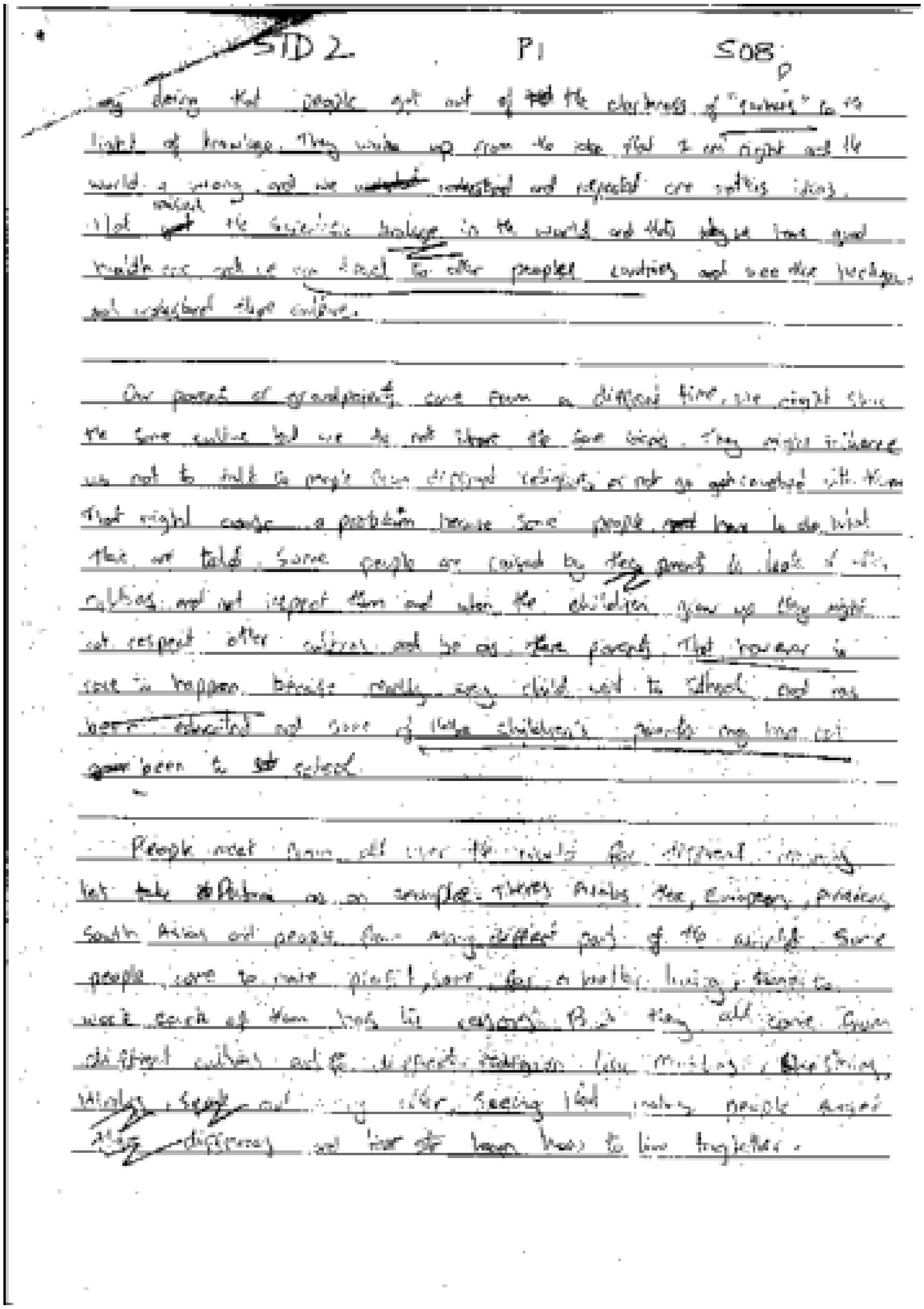

ANALYTICAL EPT CMB/PMT 2. F 07
70
60
50
40
30
20
10
0
0 10 20 30 40 50 60 70
HOLISTIC EPT/PMT 2. F 07


ANALYTICAL EPT R1/PMT 2. F 07
CRITERION EPT/PMT 2. F 07
0
10
20
30
40
50
60
70


ANALYTICAL EPT R 2/PMT 2. F 07
CRITERION EPT/ PMT 2. F 07
0
10
20
30
40
50
60
70


ANALYTICAL EPT CMB/PMT 2. F 07
CRITERION EPT/PMT 2. F 07
0
10
20
30
40
50
60
70

Appendix H
(Student Samples for Prompt 1 in Spring 2008)

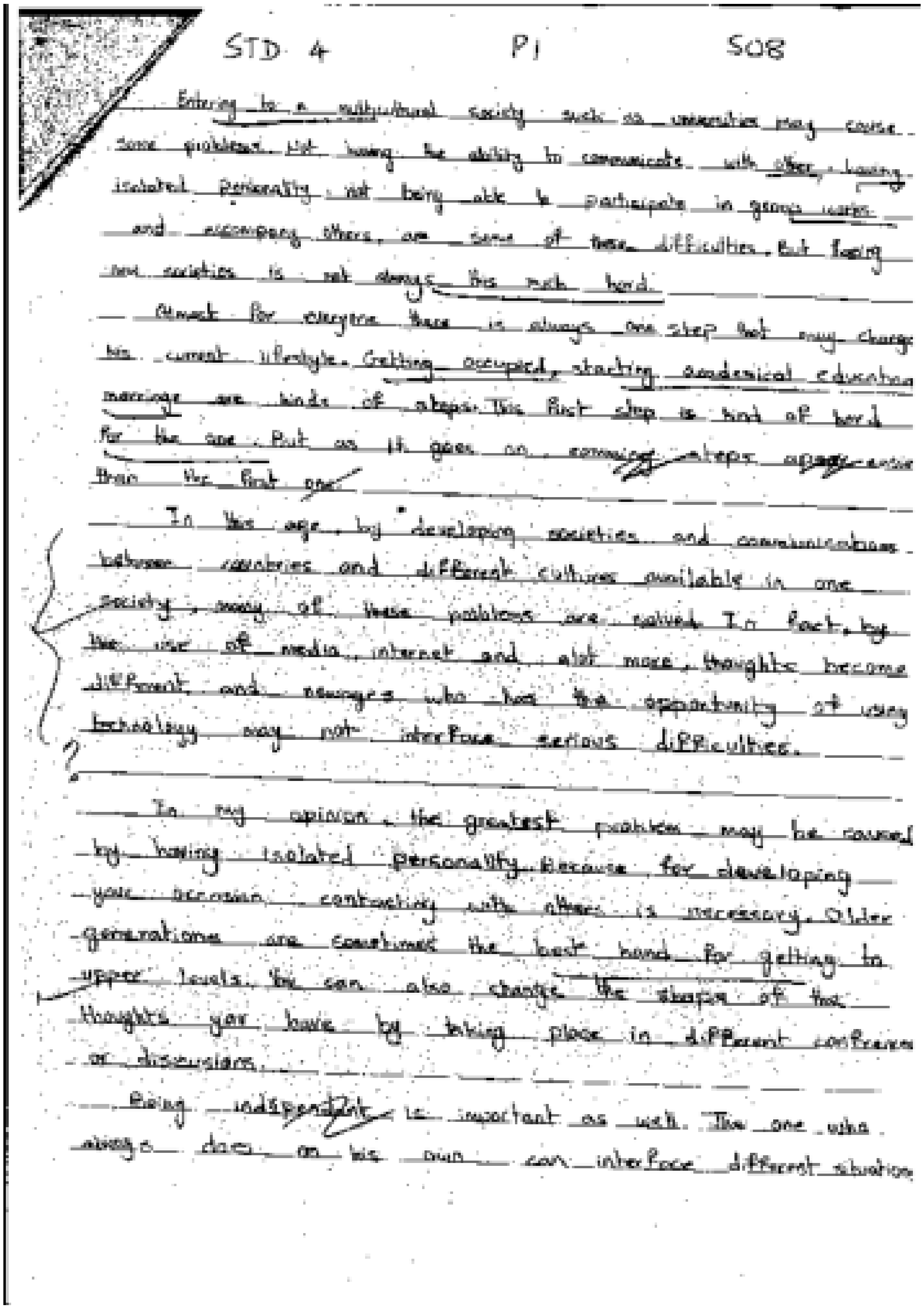


SID 2 P1 S08

ing doing that people get out of ~~[illegible]~~ the darkness of "[illegible]" to the light of knowledge. They wake up from the idea that I am right and the world is wrong, and we ~~[illegible]~~ understood and respected each others ideas. that raised ~~[illegible]~~ the scientific [illegible] in the world and that why we have good [illegible] and we can travel to other peoples countries and see their heritage and understand their culture.

Our parents or grandparents come from a different time, we might share the same culture but we do not share the same things. They might influence us not to talk to people from different religions or not to get involved with them. That might cause a problem because some people ~~[illegible]~~ have to do what they are told. Some people are raised by their parents to look at other cultures and not respect them and when the children grow up they might not respect other cultures and be as their parents. That however is rare to happen because nearly every child went to school and was been educated and some of these children's parents may have not ~~[illegible]~~ been to ~~[illegible]~~ school.

People moved from all over the world for different reasons let take ~~[illegible]~~ Britain as an example. There's Arabs, [illegible], European, African, South Asian and people from many different part of the world. Some people come to make profit, some for a better living, [illegible] to work each of them has his reasons. But they all come from different cultures and ~~[illegible]~~ different religion like Muslims, Christians, Hindus, Sikhs and many other. Seeing that makes people accept ~~[illegible]~~ differences and learn ~~[illegible]~~ how to live together.



STD 4 P1 508

Entering to a multicultural society such as universities may cause some problems. Not having the ability to communicate with other, having isolated personality, not being able to participate in group works and accompany others, are some of these difficulties. But facing new societies is not always this much hard.

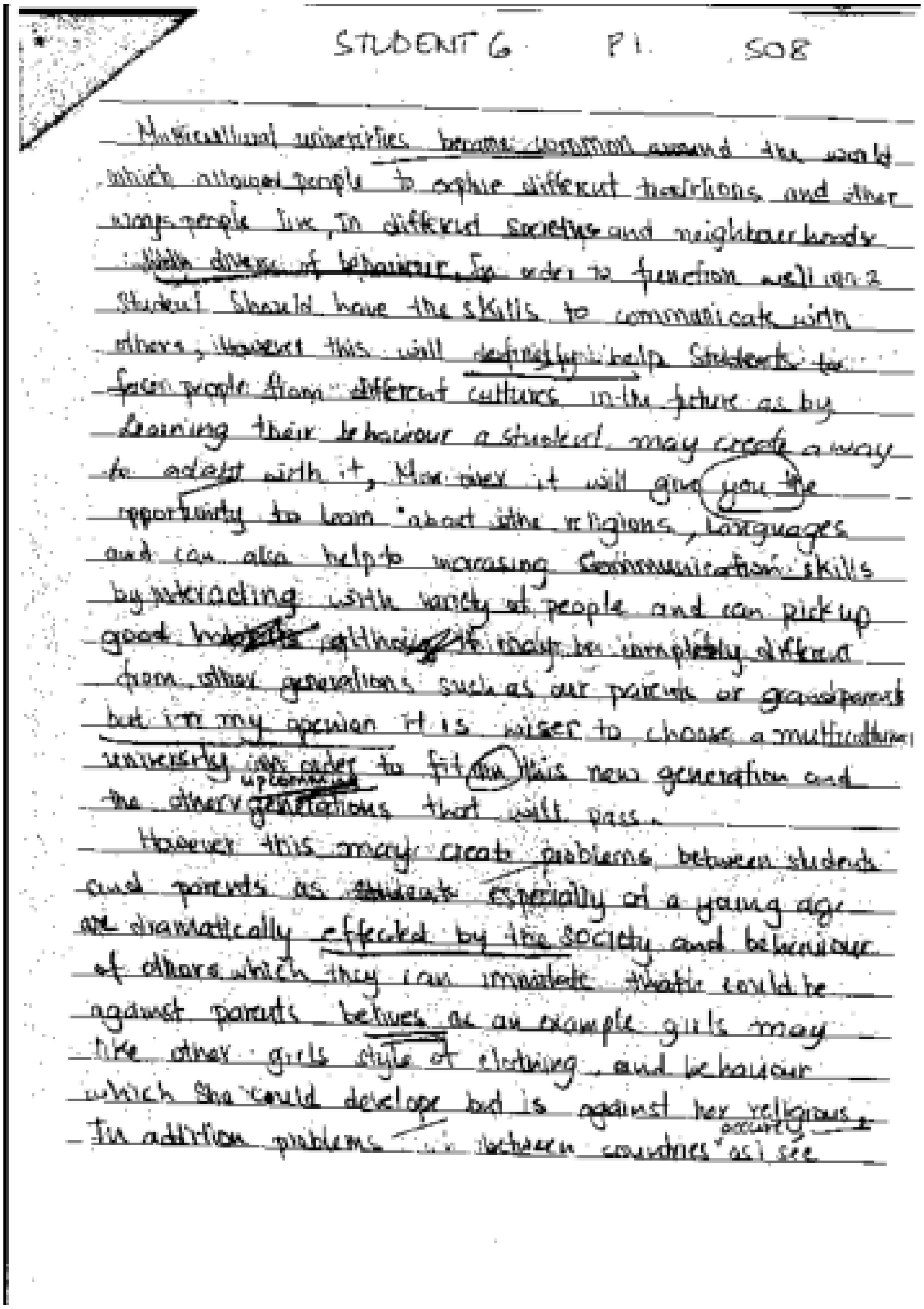
Almost for everyone there is always one step that may change his current lifestyle. Getting occupied, starting academical education, marriage are kinds of steps. This first step is kind of hard for the one. But as it goes on, comming steps ~~appear~~ easier than the first one.

In this age, by developing societies and communications between countries and different cultures available in one society, many of these problems are solved. In fact, by the use of media, internet and alot more, thoughts become different and teenagers who has the opportunity of using technology may not interface serious difficulties.

?

In my opinion, the greatest problem may be caused by having isolated personality. Because for developing your decision contacting with others is necessary. Older generations are sometimes the best hand for getting to upper levels. You can also change the shape of the thoughts you have by taking place in different conferences or discussions.

Being independent is important as well. The one who always does on his own can interface different situations



STUDENT 6 P1. S08

Multicultural universities became common around the world
which allowed people to explore different traditions and other
ways people live, In different societies and neighbourhoods
with diverse of behaviour. So in order to function well in a
Student should have the skills to communicate with
others, however this will definitely help Students to
face people from different cultures in the future as by
learning their behaviour a student may create a way
to adapt with it, Moreover it will give you the
opportunity to learn about other religions, languages
and can also help to increasing Communication skills
by interacting with variety of people and can pick up
good ~~habits~~ anything ~~which~~ it may be completely different
from other generations such as our parents or grandparents
but in my opinion it is wiser to choose a multicultural
university in order to fit ~~in~~ this new generation and
the other upcoming generations that will pass.

However this may create problems between students
and parents as ~~students~~ especially at a young age
are dramatically effected by the society and behaviour
of others which they can imitate that it could be
against parents believes as an example girls may
like other girls style of clothing and behaviour
which she could develope but is against her religious.
In addition problems between countries security as I see

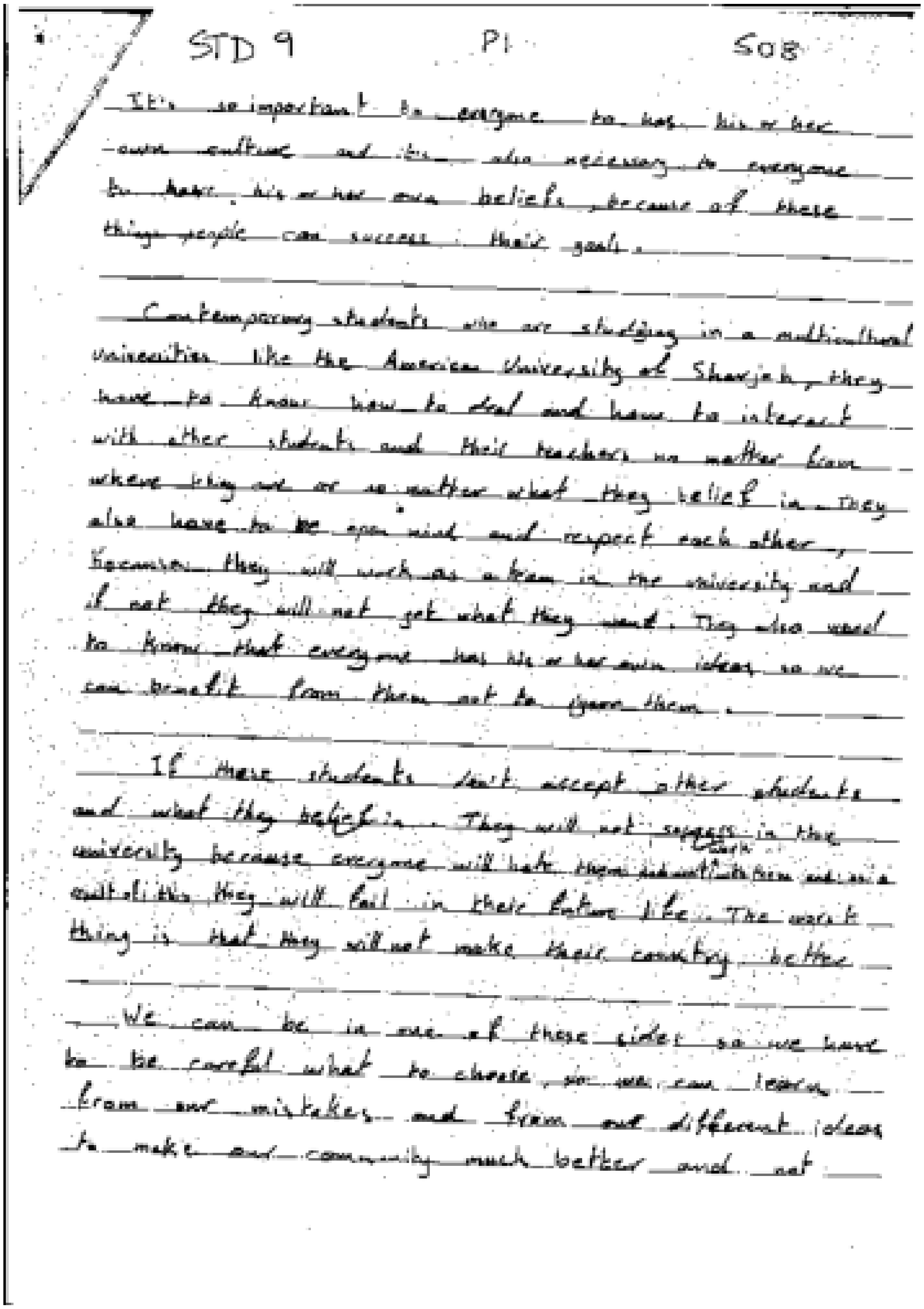
" A path to a new world"

STUDENT 6 P1 S08

Time passes and a new generation comes along, filled with youth of different lifestyle, behavior towards tradition and new experiences in life. Adolescents nowadays go through many things in life and inorder to survive this world where in, one has to accept new things. So students in multicultural universities should try to understand one another in different way.

First, communication is a key to a new path of different cultures. Students in multicultural places should try to know one another and learn the traditions of the different countries in the world and learn to respect it. When I was in a boarding school last summer in South of France I met different nationalities and when I got to know them I gained and learned many things and indeed those memories I had made me a better person.

Second knowledge also has big impact in ones life. It could be gained



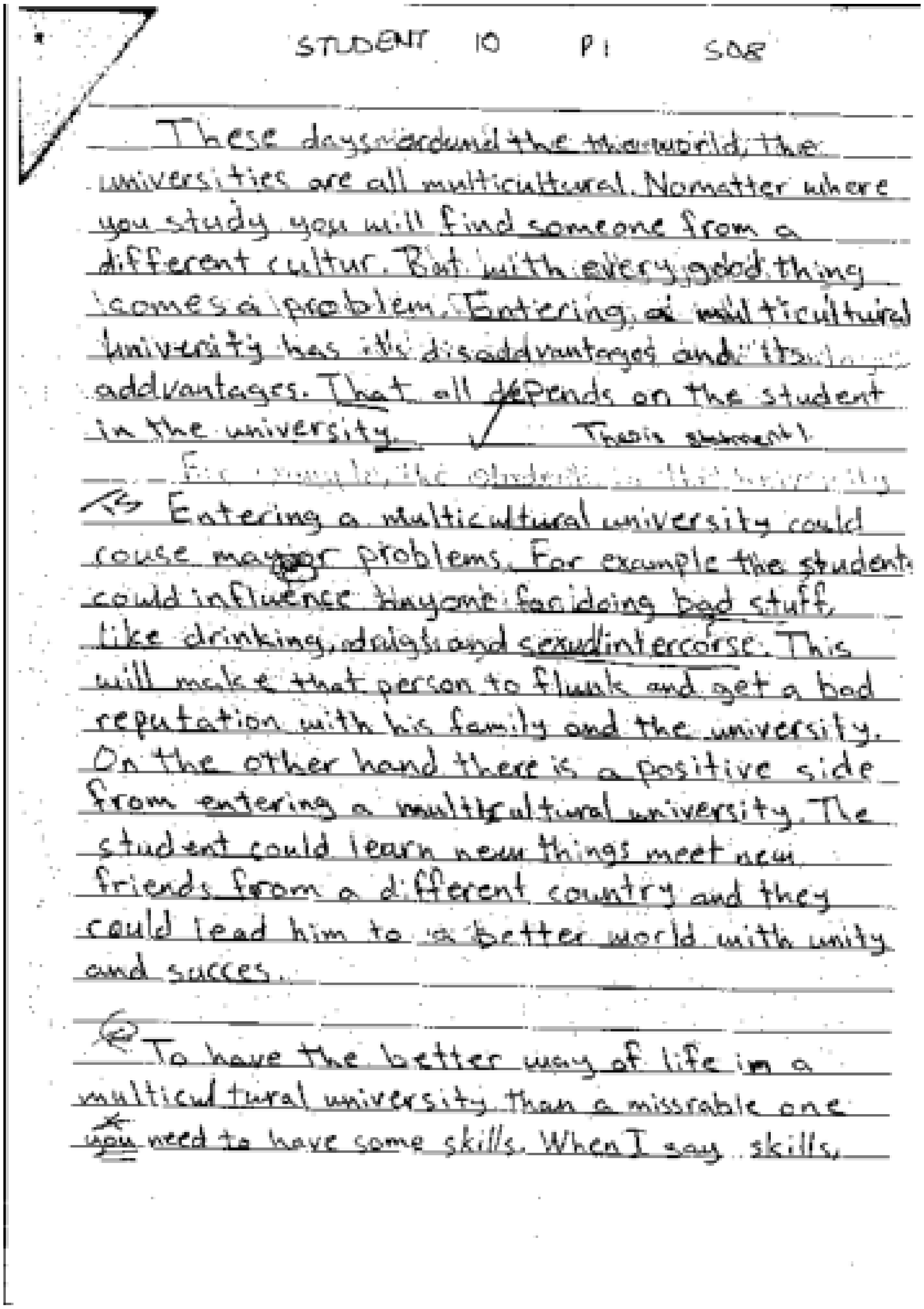
STD 9 P1 508

It's so important to everyone to has his or her own culture and it's also necessary to everyone to have his or her own beliefs, because of these things people can success their goals.

Contemporary students who are studying in a multicultural universities like the American University of Sharjah, they have to know how to deal and how to interact with other students and their teachers no matter from where they are or no matter what they belief in. They also have to be open mind and respect each other, because they will work as a team in the university and if not they will not get what they want. They also need to know that everyone has his or her own ideas so we can benefit from them not to ignore them.

If these students don't accept other students and what they belief in. They will not success in the university because everyone will hate them and won't [illegible] them and as a result of this they will fail in their future life. The worst thing is that they will not make their country better

We can be in one of these sides so we have to be careful what to choose, so we can learn from our mistakes and from our different ideas to make our community much better and not



STUDENT 10 P1 S08

These days around the ~~the~~ world, the universities are all multicultural. Nomatter where you study you will find someone from a different cultur. But with every good thing comes a problem. Entering ~~a~~ multicultural university has it's disaddvantages and its addvantages. That all depends on the student in the university. ✓ Thesis statement!

~~For example, the students in this university~~

TS Entering a multicultural university could couse mayor problems. For example the students could influence ~~anyone~~ for doing bad stuff, like drinking, drugs and sexual intercorse. This will make that person to flunk and get a bad reputation with his family and the university. On the other hand there is a positive side from entering a multicultural university. The student could learn new things meet new friends from a different country and they could lead him to a better world with unity and succes.

② To have the better way of life im a multicultural university than a missrable one you need to have some skills. When I say skills, STUDENT 11 P1 S08

The cultural complexities

Today, we live with diffrent kind of people who come from diffrent cultures, therfore we have to go with them and try to understand their culture to avoid the misundarstand that maly happen, if we did not understand them and their culture well. There are alot of cultural complexities face Students when they enter multicultural university. I'm a student in American University of Sharjah for this reason I have very good experince to have function well within this culturally diverse community. There are three main skills will help any new student who enter multicultural university like American University of Sharj.

TS Firstofall, do not be shy, try to discover other culture to be familiar with their traditinal. For example, try to read books and look into the internet pages to know much about anothe culture to have good relationship between you and anothe student who come from another culture. From my own experiance, when I entered American University of Sharjah, I fall in love with an iranian girl, so I brang yellow flower for her and I shocked when she threw the flower to my face. After that I read about iranian culture and I discoverd that in their culture when you bring yellow flower that mean you hate her ⓔ As a result I went to her and explaned that for here. In my opinion, if you discover another culture, you will have good relationship with other people from diffrent culture.

In addition, you have to give great respect to another people who come from diffrent culture than you even if you see their culture is wrong. As you are a student in multicultural university, try to respect all student and have their heart in your head. For example, when I was

STD 14 P1 S08

got advantages in dealing with people such as easy going and live a simple life. This consider helpful to anyone who wants to live a simple life. Also our culture (Arabic culture) got good traits like generosity and loelty.

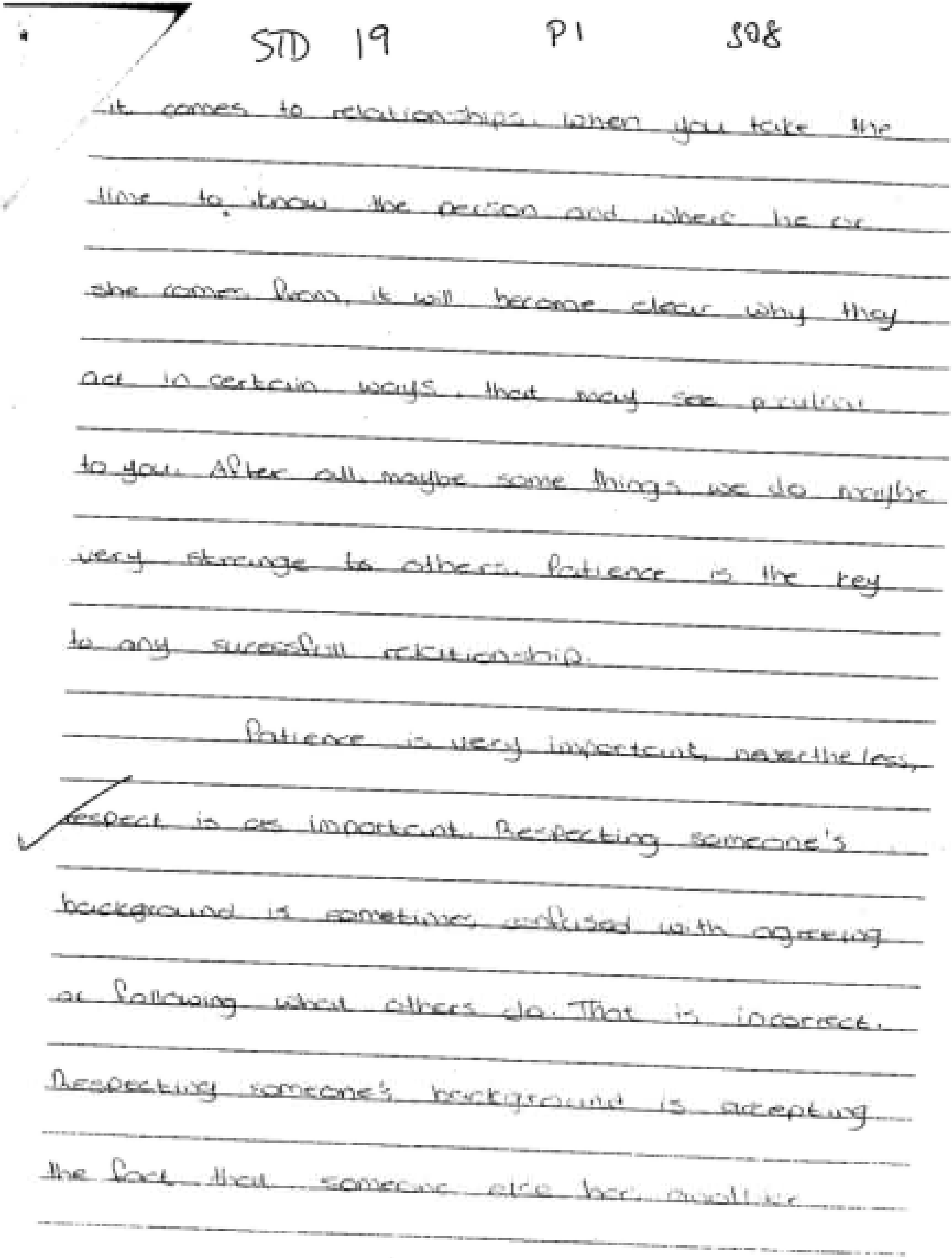
Finally, is solving problems in diffrent ways. Not All clutures are the same people got different ways in thinking and that's derived from thier own culture. There defferace coulld help solving problems. That goes with the example of mine when I was a studat in London. At that time

STUDENT 18 P1 S08

Cultural Complexities

Being in a good university is what most of students aim for. Many students travel abrod to join well-known university, which they dream to join. The fact is that when they reach to that university, they face cultural complexities. I think the most complexities that may young people face there are to have difficulties to communicate with others and to be logly. Thesis statement

Firstofall, Communicate with others can be very hard for the forighn students because they come from different cultures; forexample, Once I met a forighn Student from Kongo at AUS, which is a multicultural university, with whome I was trying to communicate, yet

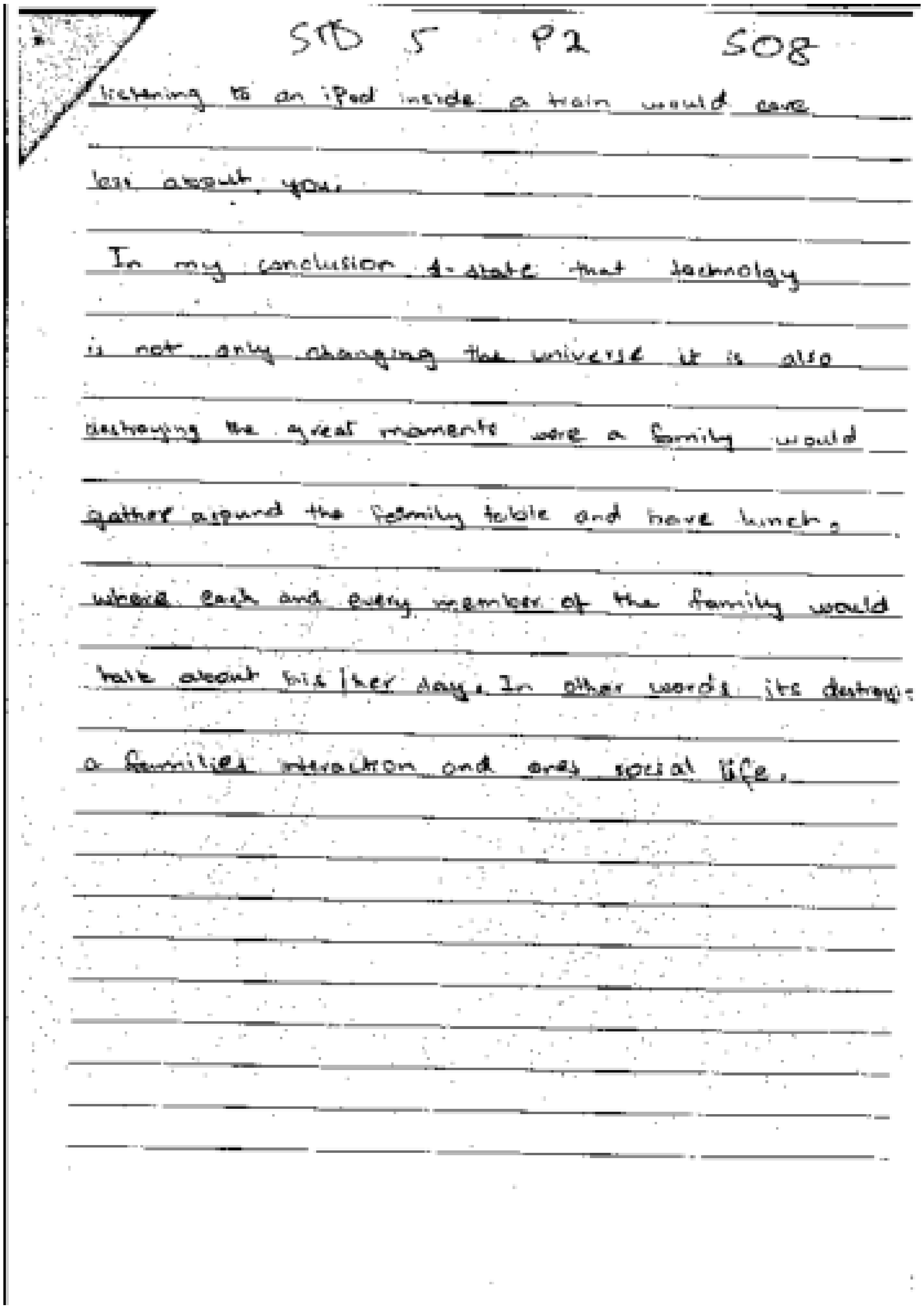


STD 19 P1 S08

it comes to relationships. When you take the time to know the person and where he or she comes from, it will become clear why they act in certain ways, that may see peculiar to you. After all maybe some things we do maybe very strange to others. Patience is the key to any successfull relationship.

Patience is very important, nevertheless, respect is as important. Respecting someone's background is sometimes confused with agreeing or following what others do. That is incorrect. Respecting someone's background is accepting the fact that someone else has another

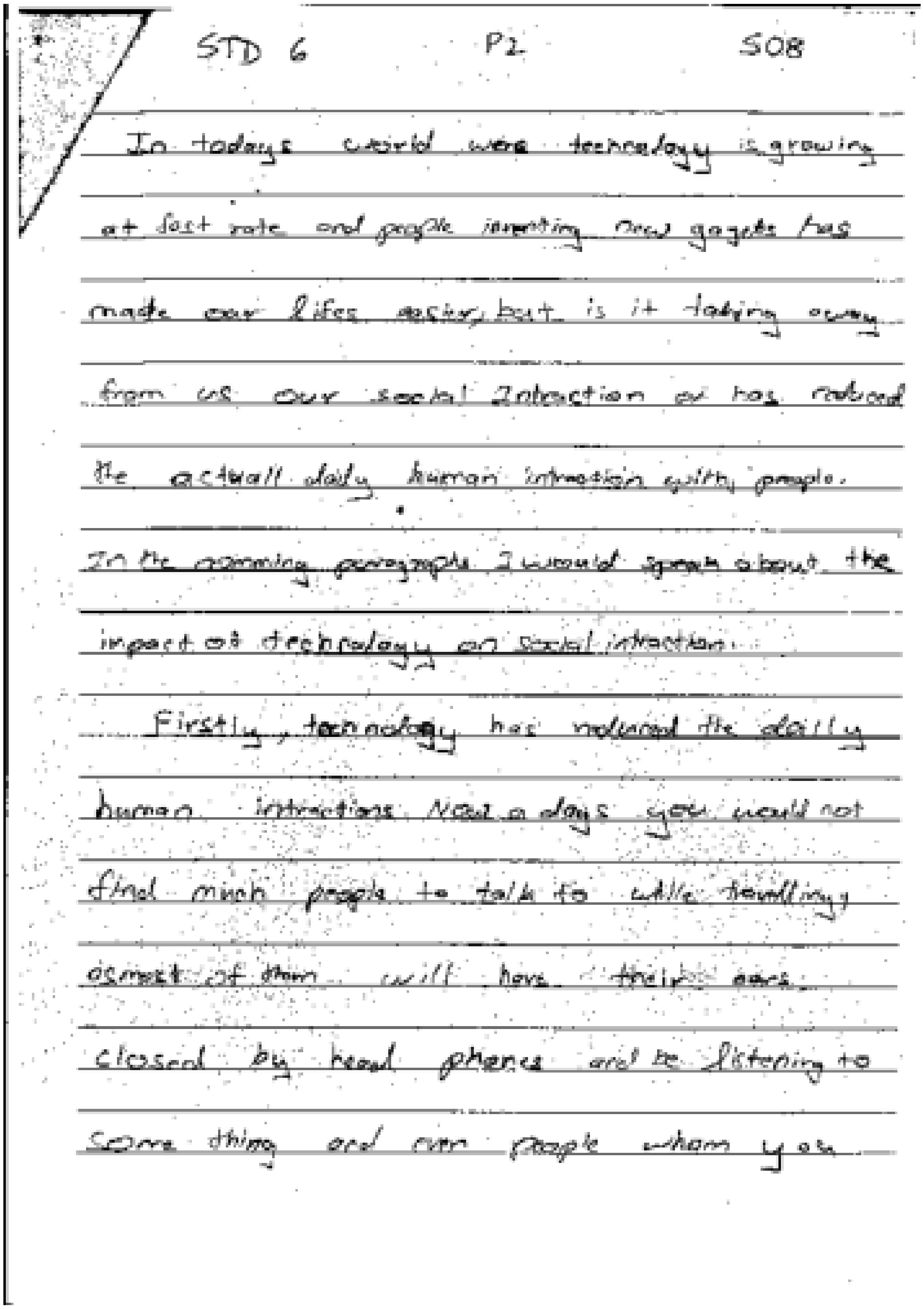
Appendix I
(Student Samples for Prompt 2 in Spring 2008)



STD 5 P2 S08

listening to an iPod inside a train would care less about you.

In my conclusion I state that technolgy is not only changing the universe it is also destroying the great moments were a family would gather around the family table and have lunch, where each and every member of the family would talk about his/her day. In other words its destroying a families interaction and ones social life.



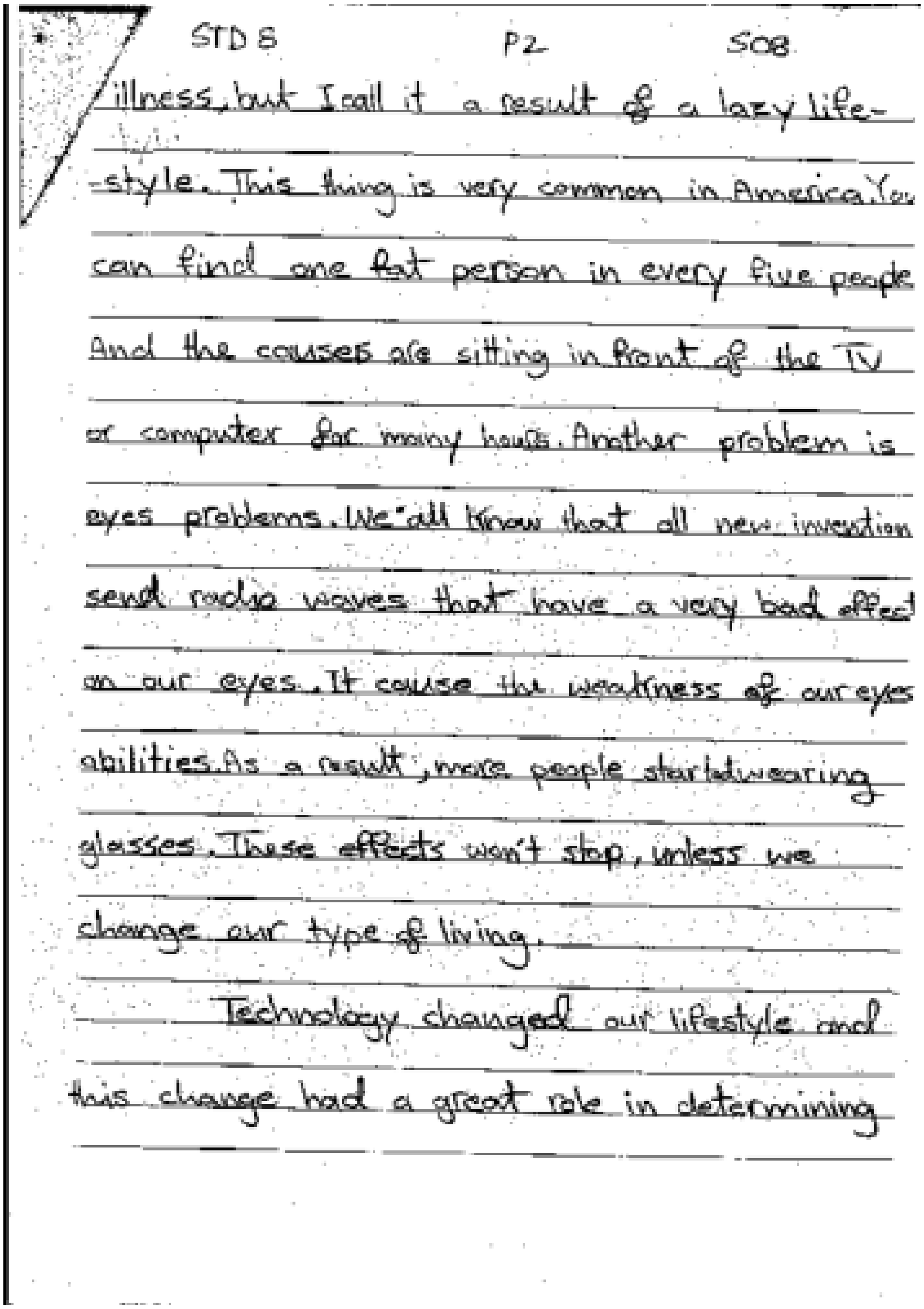
STD 6 P2 S08

In todays world were technology is growing at fast rate and people inventing new gagets has made our lifes easier, but is it taking away from us our social Intraction or has reduced the actuall daily human intraction with people. In the comming paragraphs I would speak about the impact of technology on social intraction.

Firstly, technology has reduced the daily human intractions Now a days you would not find much people to talk to while travelling asmost of them will have their ears closed by head phones and be listening to some thing and even people whom you

STD 7 P2 S08

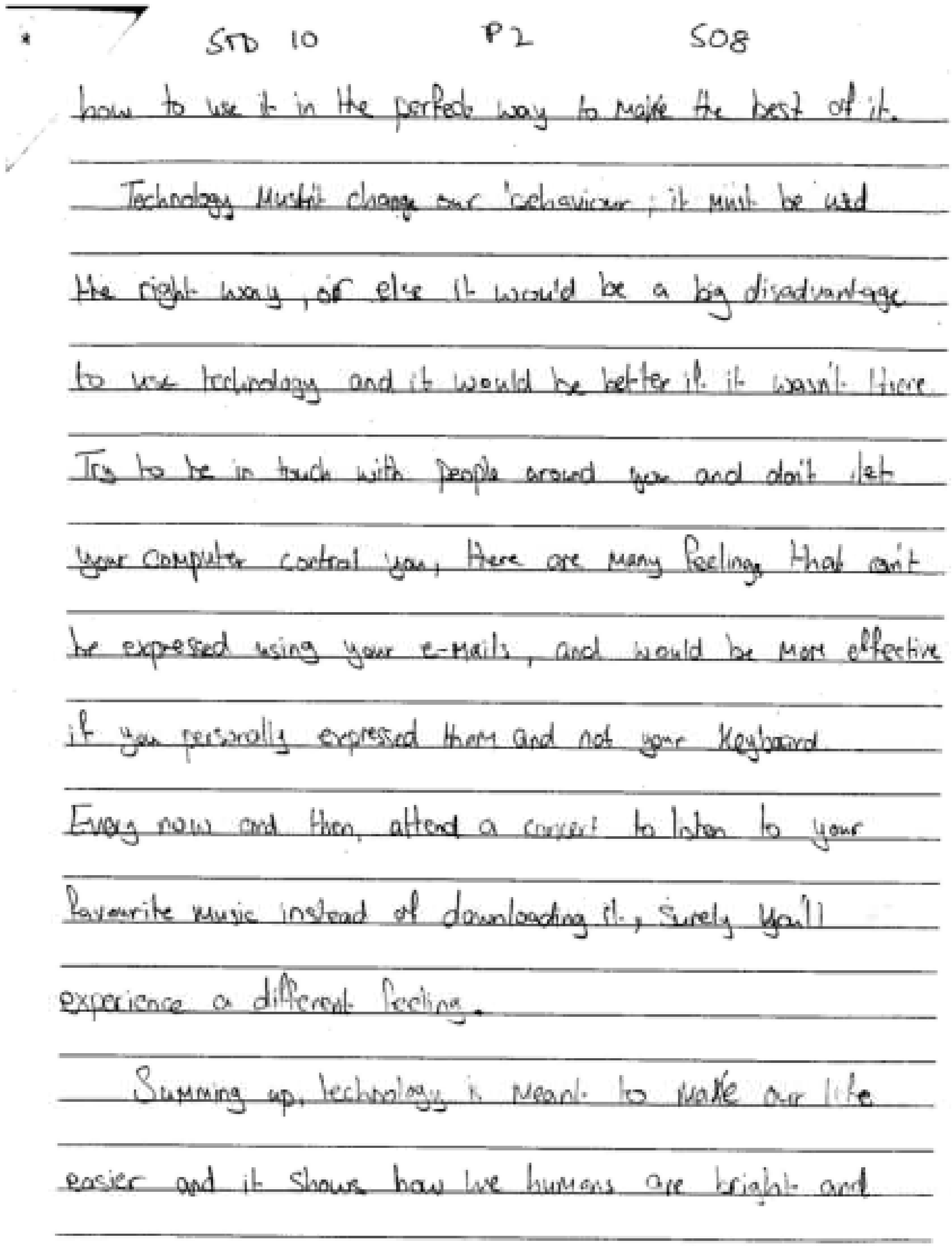
There are many facts that changed the world commonucation nowadays. Never the less people used to commonucate by letters and meeting only but with the development that has changed the social interaction people thesedays are using many ways to commonucate with each other, as an example: Cell phones, internet and many other things. In addition people behavior changed since the technology has developed in many ways like all people used to write alot but these days afew people write and not that much. Some people say it is growing the people minu and some people say it is a waste of human minds. It is a fact that people used to look for information for a long time as going from place to another to find this information and some used to travel, but nowadays people sit infront of the computer's monitor and just severe the net looking for it and this won't take time as it used to take before. The most shocking thing that is when people are having many friends nowadays in no time and just without meeting



STD 5 P2 S08

illness, but I call it a result of a lazy lifestyle. This thing is very common in America. You can find one fat person in every five people And the causes are sitting in front of the TV or computer for many hours. Another problem is eyes problems. We all know that all new invention send radio waves that have a very bad effect on our eyes. It cause the weakness of our eyes abilities. As a result, more people started wearing glasses. These effects won't stop, unless we change our type of living.

Technology changed our lifestyle and this change had a great role in determining



STD 10 P2 S08

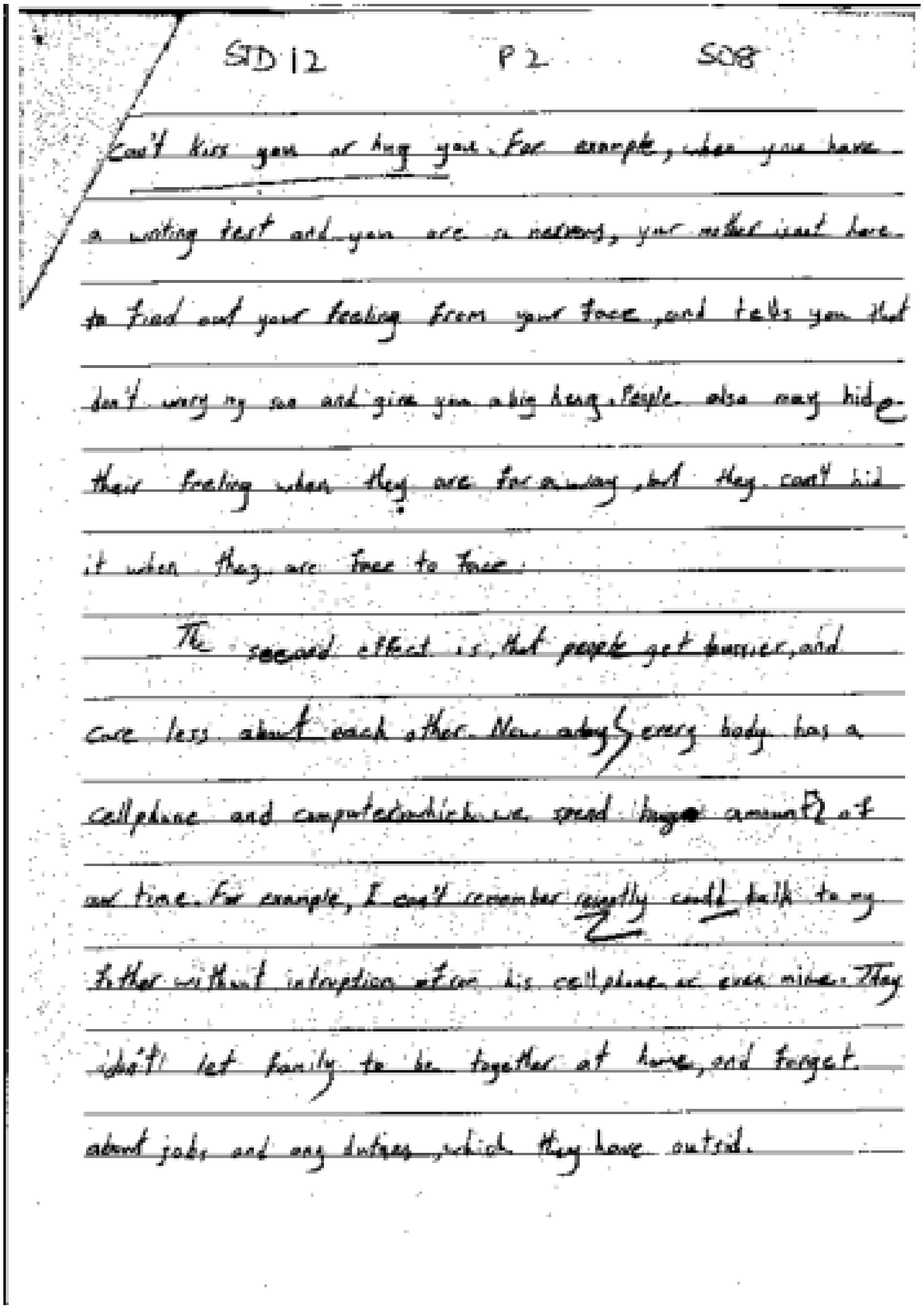
how to use it in the perfect way to make the best of it.

Technology Mustn't change our behaviour; it must be used the right way, or else it would be a big disadvantage to use technology and it would be better if it wasn't there. Try to be in touch with people around you and don't let your computer control you, there are many feeling that can't be expresed using your e-mails, and would be more effective if you personally expressed them and not your Keyboard. Every now and then, attend a concert to listen to your favourite music instead of downloading it, Surely you'll experience a different feeling.

Summing up, technology is meant to make our life easier and it shows how we humans are bright and

STD 11 P2 S08

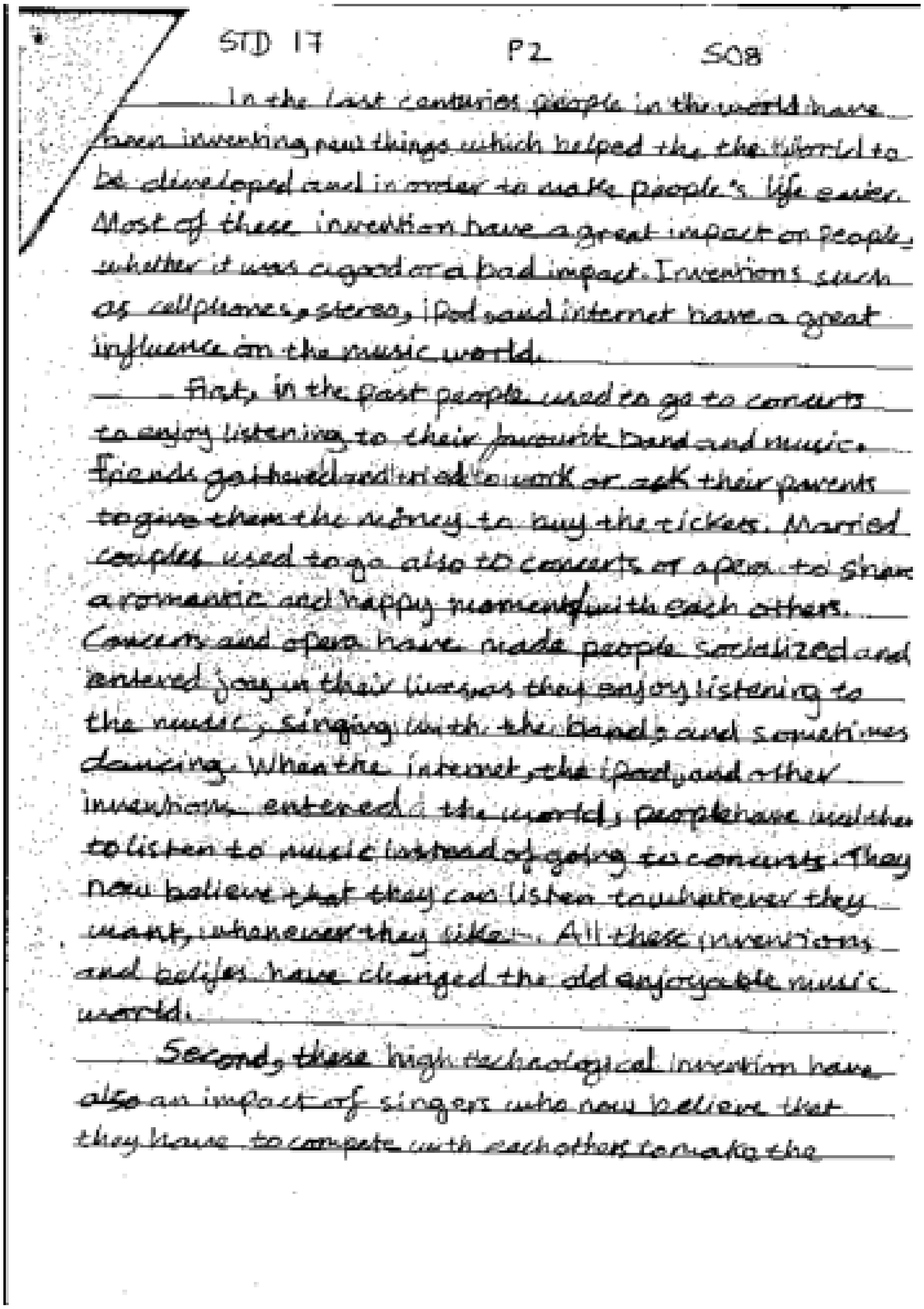
while they are listening, they can catch many good ideas which is related to their behaviour. Consequently, they act them among each other. In addition, almost all people use the world wide web in order to increase their Knowledge in different fields. For example, people can read distinct things whether to enjoy themsalves or to increase their Knowledge, so they can have information about behaviours, So that it can help them when dealing with people. In fact, not only old people can get the advantages of the internet, but also Kids can. This an example



STD 12 P2 S08

can't kiss you or hug you. For example, when you have a writing test and you are a nervous, your mother isn't here to find out your feeling from your face, and tells you that don't worry my son and give you a big hug. People also may hide their feeling when they are far away, but they can't hid it when they are face to face.

The second effect is, that people get busier, and care less about each other. Now adays every body has a cellphone and computer which we spend huge amounts of our time. For example, I can't remember exactly could talk to my father without intruption of from his cellphone or even mine. They didn't let family to be together at home, and forget about jobs and any duties, which they have outsid.

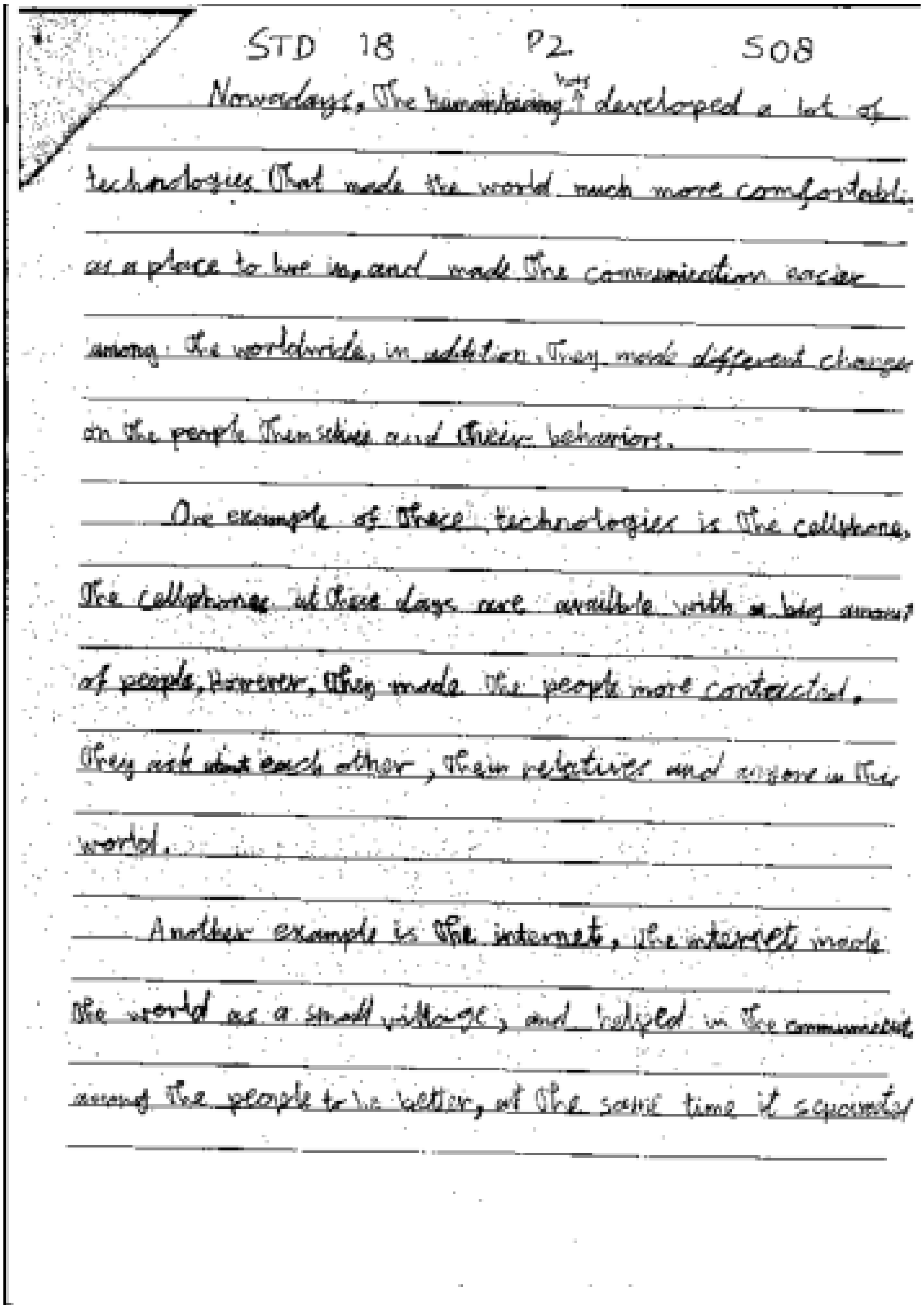


STD 17 P2 S08

In the last centuries people in the world have been inventing new things which helped the the World to be developed and in order to make people's life easier. Most of these invention have a great impact on people, whether it was a good or a bad impact. Inventions such as cellphones, stereo, iPod and internet have a great influence on the music world.

First, in the past people used to go to concerts to enjoy listening to their favourite band and music. Friends gathered and tried to work or ask their parents to give them the money to buy the tickets. Married couples used to go also to concerts or opera to share a romantic and happy moments with each others. Concerts and opera have made people socialized and entered joy in their lives as they enjoy listening to the music, singing with the bands and sometimes dancing. When the internet, the iPod, and other inventions entered the world, people have used them to listen to music instead of going to concerts. They now believe that they can listen to whatever they want, whenever they like. All these inventions and belifes have changed the old enjoyable music world.

Second, these high technological invention have also an impact of singers who now believe that they have to compete with each others to make the

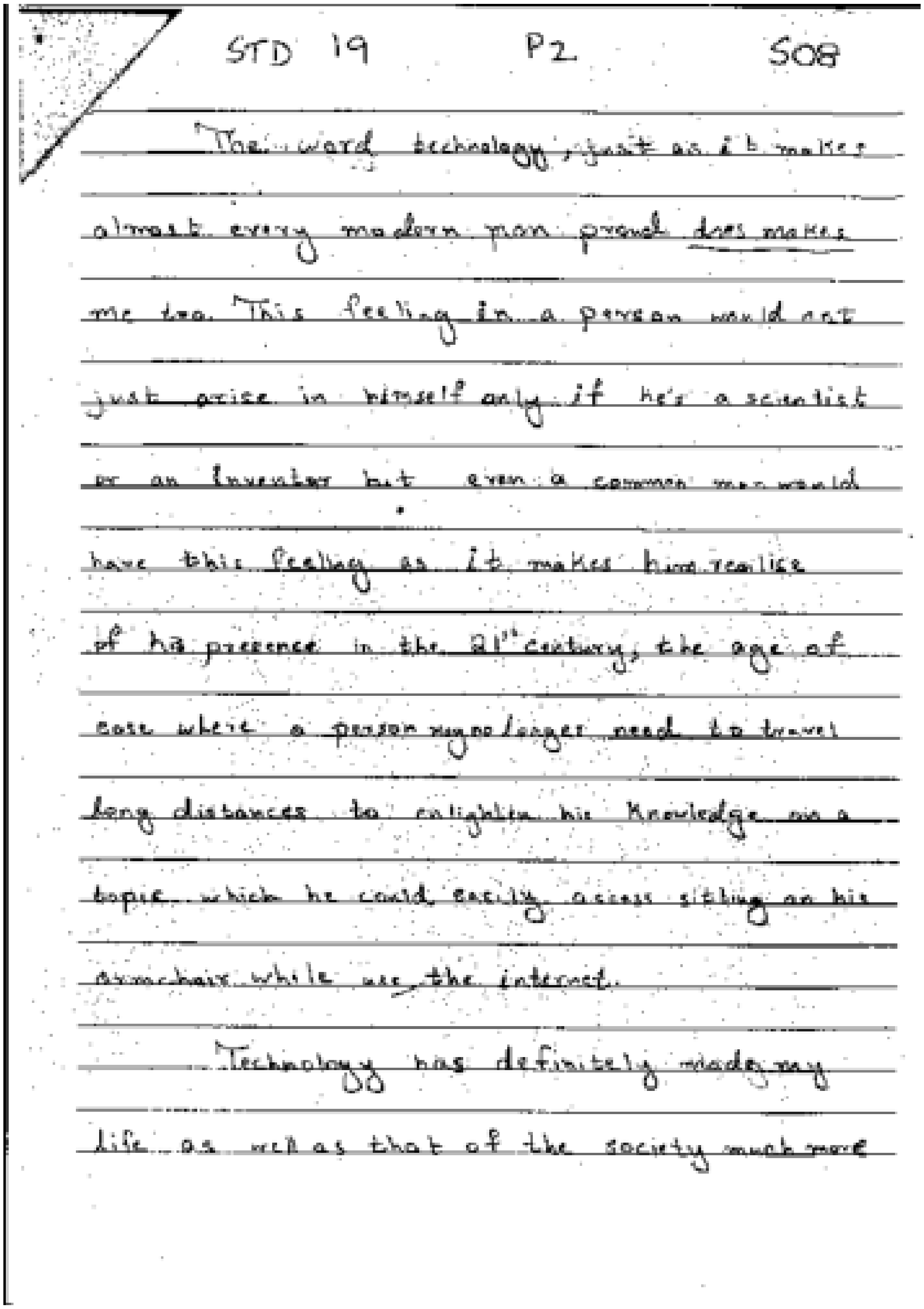


STD 18 P2 508

Nowadays, The humanbeing has developed a lot of technologies That made the world much more comfortable as a place to live in, and made The communication easier among the worldwide, in addition, They made different changes on the people Themselves and Their behaviors.

One example of These technologies is The cellphone, The cellphones at these days are available with a big amount of people, however, they made the people more contacted. They ask about each other, Their relatives and anyone in The world.

Another example is The internet, The internet made The world as a small village, and helped in The communication among The people to be better, at The same time it separates



STD 19 P2 S08

The word technology, just as it makes almost every modern man proud does makes me too. This feeling in a person would not just arise in himself only if he's a scientist or an Inventor but even a common man would have this feeling as it makes him realise of his presence in the 21st century, the age of ease where a person no longer need to travel long distances to enlighten his Knowledge on a topic which he could easily access sitting on his armchair while use the internet.

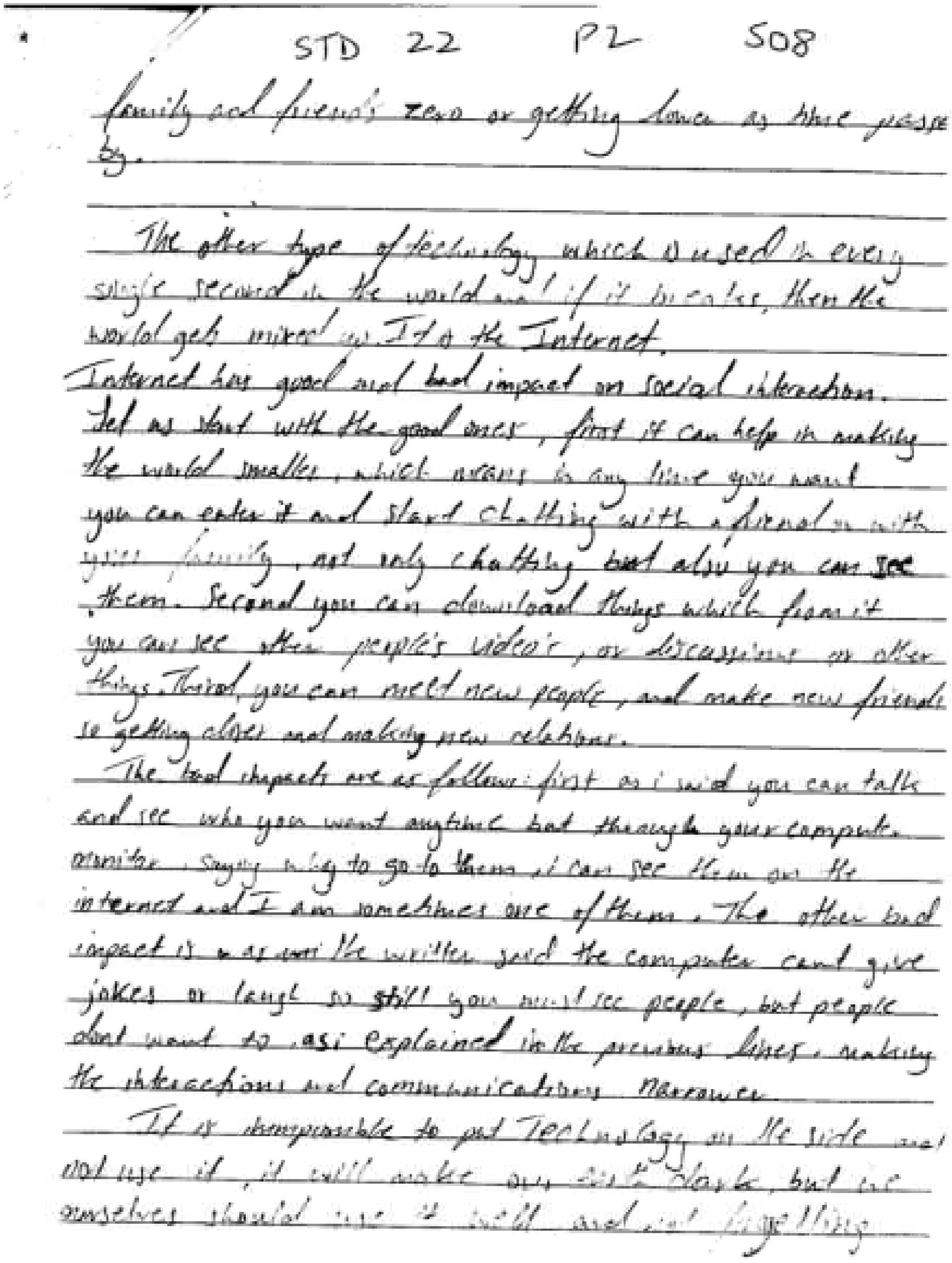
Technology has definitely made my life as well as that of the society much more

STD 20 P2 S08

Nowadays, for us its almost impossible to imagine the life, without cars, cellphones or and the television. Technology has affected many different factors of our lives, one of which is social factor. There are alot of different people, who depend on, and adicted to the internet, television, cell phones and other technologies.

The internet has a lot fuctions that people can use. Eversince humanity invented the internet, people decided to use it as a communication system. At the begining people were using it secretly, but when it was exposed to social live, people all over the world starged to use it. Nowadays internet is so comon in our world that its almost impossible not to know about it. But "with every advantage comes a disadvantage". There are a lot of people, especialy teenagers, who are adicted to internet games and chats and who could spend hours surfing the internet.

We can spend hours talking to our frinds by using cellphones or we can spend hours watching favorite TV show. from my point of view, technology has ruined social life. for example, Many people prefer to communicate with each other by cell phones, when they can get together and socialize.



STD 22 P2 508

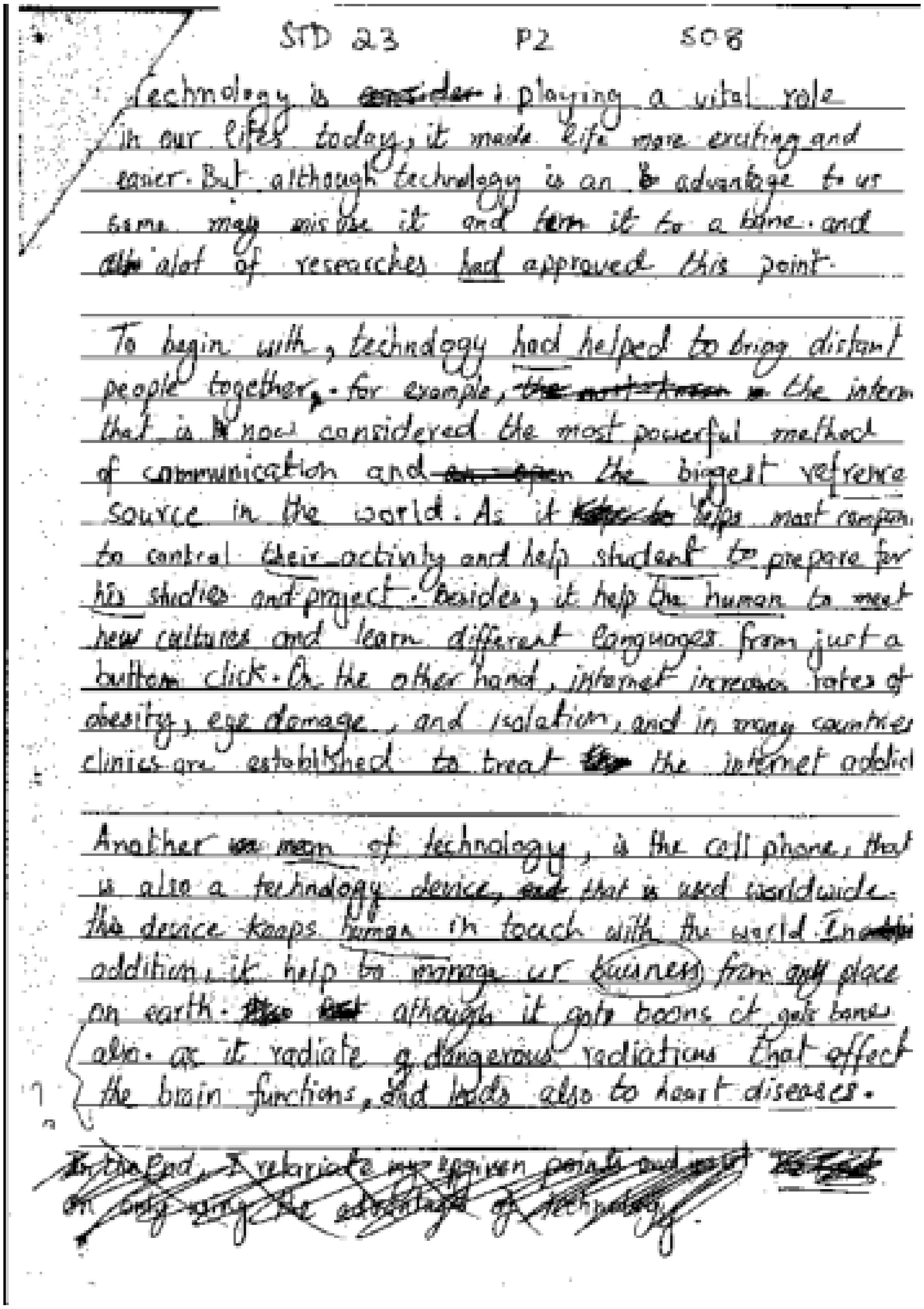
family and friend's zero or getting lower as time passe by.

The other type of technology which is used in every single second in the world and if it breaks, then the world gets mixed up. It is the Internet.

Internet has good and bad impact on social interaction. Let us start with the good ones, first it can help in making the world smaller, which means in any time you want you can enter it and start chatting with a friend or with your family, not only chatting but also you can see them. Second you can download things which from it you can see other people's video's, or discussions or other things. Third, you can meet new people, and make new friends so getting closer and making new relations.

The bad impacts are as follows: first as i said you can talk and see who you want anytime but through your computer monitor. Saying why to go to them i can see them on the internet and I am sometimes one of them. The other bad impact is as the writer said the computer cant give jokes or laugh so still you must see people, but people dont want to as i explained in the previous lines, making the interactions and communications narrower.

It is impossible to put Technology on the side and not use it, it will make our life dark, but we ourselves should use it well and not forgetting



STD 23 P2 508

Technology is ~~consider~~ playing a vital role in our lifes today, it made life more exciting and easier. But although technology is an advantage to us some may misuse it and ~~term~~ it to a bane. and ~~also~~ alot of researches had approved this point.

To begin with, technology had helped to bring distant people together. for example, ~~the most known~~ the intern that is now considered the most powerful method of communication and ~~an open~~ the biggest refrence source in the world. As it ~~keeps to~~ helps most compan. to control their activity and help student to prepare for his studies and project. besides, it help the human to meet new cultures and learn different languages from just a buttom click. On the other hand, internet increases rates of obesity, eye damage, and isolation, and in many countries clinics are established to treat ~~the~~ the internet addic

Another ~~m~~ mean of technology, is the cell phone, that is also a technology device, ~~and~~ that is used worldwide. this device keeps human in touch with the world. In addition, it help to manage ur business from any place on earth. ~~the but~~ although it gots boons it gets bane also. as it radiate a dangerous radiations that affect the brain functions, and leads also to heart diseases.

~~In the End, I retariate my given points and insist on only using the advantages of technology.~~

VITA

Mary Anne John completed her MA Literature degree from Calcutta University, India in 1992 prior to which she completed her BEd degree from Loreto College in 1989. She started her career in a Methodist school in Calcutta in 1990 and taught grades four to twelve for five years. In 1995, she came to Abu Dhabi when her family relocated to the UAE. She first taught at high school and later on taught communicative English to adults at Emirates. While working at Emirates, she completed her CELTA in 2004 from the Dubai Women’s College and in 2005 joined the MA TESOL program at the American University of Sharjah (AUS). That same year she joined AUS Department of Writing Studies, where she teaches freshmen writing courses. Mary Anne John believes that teaching is more a vocation than a profession, and for her every day is a learning experience. Each batch of students helps her discover new facets of the teaching-learning situation and for her as a teacher, life is all about being willing to learn.