Human Factors in Holistic Assessment of Writing

A Heuristic (Pilot) Study

Swathi M. Vanniarajan
San José State University

Holistic assessment is generally the tool of assessment used in evaluating student writing samples. Though the tool enjoys a great deal of face validity and inter-rater reliability, there are not many studies that have tried to investigate how raters achieve the tremendous task of overcoming their personal element in the scoring process. The heuristic pilot study reported in this paper attempts to fill in the gap in the research literature by exploring through one-on-one introspective interviews and questionnaire-surveys how the raters react and respond to the cognitive and affective demands made on them during the holistic evaluation sessions and how their personal variables, such as their personal characteristics and biases, affect their scoring process. The results of the study indicated that in spite of the raters' experience, background, training, and the norming process given at the beginning of every holistic evaluation session as well as the tool's outward inter-rater reliability, there is indeed a myriad of differences among the raters and that not all human elements can be compromised in the evaluation process.

Holistic evaluation, as a tool of assessment, enjoys a great deal of face validity largely due to its practicality. Research also has shown that holistic evaluation has high inter-rater reliability. Experience indicates that instructors with at least 3 years of experience in both teaching and holistic assessment take about 2 to 4 minutes to assess an essay of about 250 to 400 words, whereas instructors with less experience take about 3 to 5 minutes per essay. While 2 to 5 minutes for assessing a student essay seem to be remarkably short compared to about an hour's time that a student writer takes to read the prompt, think, and write, analyses indicate that there is consistency in rater scores and a discrepancy of 2 or more score on a rating scale of 1 to 6 between two raters[1] is only about 6% at least in the setting in which this research study was conducted. If raters are able to make fast conclusions, then what kind of evaluative response do they carry out while reading student writings? How do they react to various demands made on them during the reading day? How do they deal with cognitive and affective constraints under which they have to read and rate the essays? Who,

[1] I wish to make a distinction between raters and readers. Raters for the purposes of this paper will refer only to those who read for the purposes of formal evaluation. In contrast, readers will merely refer to those who read and do not make any formal assessment of the materials they read. Also note that raters and scorers are used interchangeably throughout the paper.

in their opinion, is an ideal rater? The existing research literature on holistic assessment process, unfortunately, does not provide complete answers to all of the above questions. The reason may be because the research is yet to recognize that holistic evaluation has a second dimension, the human (rater) element on top of the scoring grid and its interpretation.

Even those studies that have attempted to touch on the human element in holistic evaluation downplay its role by claiming that training of raters during the norming process can minimize or completely negate its role during the evaluation process (Freedman 1979, Huot 1993, Shohamy, Gordon & Kraemer 1992, Weigle 1994). This is an interesting claim especially in light of how holistic evaluation as a process is defined. On the one hand, the goal of holistic evaluation is to measure the impact of the essay on the rater in terms of the criteria provided (Williamson 1993) and on the other hand, what is important is not the raters' perception of what the impact is but the group's collective opinion of how closely the essay can be matched with one of the rating scales in the criteria (Purves 1992, Williamson 1993). If this is the way in which holistic evaluation has to be carried out, then how difficult or easy the assessment process is for raters? Is it possible for raters to achieve unanimity, and thereby inter-rater reliability in the evaluation process in spite of differences in their experience, background, training, preferences, and beliefs? This study is a modest attempt to investigate the difficulties in holistically evaluating essays as perceived by raters themselves and to find out how they strive to overcome these in achieving inter-rater reliability.

This heuristic pilot study has two broad goals: to explore how the raters, using scoring criteria, react and respond to the demands of the holistic evaluation process and to describe from the raters' perspective how their personal variables affect their scoring process. The results of the study indicated that in spite of their experience, training, and the norming process given at the beginning of every holistic evaluation session, as well as the outward inter-rater reliability in holistic scoring, there was indeed a myriad of differences among the raters, and not all of the human elements could be neutralized in the evaluation process.

The paper is divided into three sections. Section 1 provides a short review of the existing literature on human elements in holistic assessment. Section 2 describes the research study and its major findings. Section 3 lists the limitations of the study and provides suggestions for further

research in this area.

I

Review of the Literature on Human Elements in Holistic Evaluation

Human factors in holistic assessment

Rater reliability in holistic assessment process can be defined as consistency in one's scoring process for essays of the same quality at any point in the scoring process as well as consistency of agreement among raters (Huot 1990, Legg 1998). It is intriguing to note here that consistency in one's scoring process (intra-rater reliability) itself is a prerequisite for attaining inter-rater reliability. While intra-rater reliability can be negatively affected by a variety of reasons related to the human elements in the scoring process such as the tiring nature of the scoring process, individual factors, and cognitive and affective constraints, inter-rater reliability can be affected by the factors that raters do not share the same background or personal beliefs. (Pula & Huot 1993). What follows is a description of how human factors affect one=s intra- and inter-rater reliability in holistic evaluation.

Factors that may affect intra-rater reliability in holistic evaluation

Holistic evaluation is a tiring process and psychologically debilitating for various reasons. It is self-evident that raters need to have a sustained focus on the process of scoring for a long period of time. It is important to note here that in many settings, it takes the entire day. As such, raters generally find the task of staying focused while scoring papers highly demanding and psychologically debilitating. Wolcott (1998) points out that raters may over reward or penalize an essay when their attention begins to wander off.

In addition to tiring sessions, the raters also have to read the essays written in response to the same prompt one after another for long periods of time with very few resting sessions; such a process can also turn out to be a tedious, drudgerous, and uninteresting experience, especially if the prompt has been perceived to be unchallenging and non-controversial by some raters. In such a situation, it is possible that raters may find it very difficult to discriminate between essays since all essays may seem alike.

Yet another human element that may be a factor in intra-rater consistency is the speed with which raters read essays. If some raters happen to be readers who wish to make decisions only after

thoroughly reading student essays, then this also can lead to boredom and fatigue while scoring, coupled with feelings of anxiety about not being on par with the number of essays read by other raters. Then, they may tend to sacrifice accuracy for efficiency.

Factors that may affect inter-rater reliability in holistic evaluation

Research indicates that inter-rater reliability in holistic evaluation may get affected by raters' preferences, training, and background.

Not all student essays are alike; they do differ from one another for a variety of reasons. For example, some essays may contain excellent content but may be riddled with errors in structure and word usage; alternately, some essays may contain few mechanical errors and yet be shallow and simplistic in their treatment of content. According to Wolcott (1988), "such papers are always difficult to score, and one reader may reward the content, while the other reader may penalize the same paper for its problems, with the result that discrepant scores arise. Or a paper may take a truly unconventional approach, which is rewarded by one reader and penalized by another" (p. 80). Studies by Barritt, Stock, & Clark (1986) and Freedman (1984) indicated that the raters' self-decided criteria as to what an exit level student essay should look like negatively affected the inter-rater reliability. Studies by Erickson (2001), Charney (1984), and Stach (1987), furthermore, indicated that scores given by various raters did not correlate well with the number of spelling errors, mechanics, vocabulary, essay length, syntactic complexity, the length of the syntactic structures, or even the appearance of an essay.

Rater reliability has been shown to be at risk by the rater background also. For example, discrepancy in scores seems to be on the higher side if the raters happen to be a mix of both native and nonnative speakers, and are collectively evaluating essays written by both native and nonnative speakers. Zhang (2000) found that the essays written by nonnative speakers were evaluated differently by native and nonnative raters and that the difference was significant, especially when the nonnative rater belonged to the ethnic group of the nonnative writer. It is interesting to note in this regard that Huot (1993) reported in his research study that untrained raters attended to the structural aspects of the writing rather than to the communicative aspects.

Rater errors can also occur due to a variety of cognitive and affective factors. According to Wolcott (1998), errors based on cognitive factors could fall into three types: (1) halo errors, (2) leniency errors, and (3) range restriction errors. Halo errors occur when a rater reads an essay of average quality after reading an excellent essay. The average essay seems to be of very poor quality by comparison and the rater may err by giving it a failing score. Leniency errors reflect raters' nature to be either strict or lenient in their evaluation process. Range restriction errors occur when raters suffer from a tendency to avoid giving essays either low or high scores so as not to end up with discrepant scores.

The challenge

The review of the literature reveals that though the human factors in the holistic scoring process have been well understood, whether these factors actually affect the scoring process has not been studied very well. This may be due to three factors. First, language testing researchers are more concerned with the psychometric models of assessment. Second, they think that they need to be concerned only with the statistical issues related to reliability and the validity of the instruments they use. And third, they have yet to recognize that the holistic scoring process is actually a process in which a great deal of human element is involved. In order to know the inherent validity issues in holistic assessment, it is important that we initially understand how raters successfully negotiate with the cognitive and affective constraints that they have to overcome in the scoring process. In sum, this paper makes an attempt to understand from the raters' perspective how in their opinion they meet the demands made on them and are able to overcome the propensity to make errors in the evaluation process..

II

Research Study

The Study

The study was conducted at a reading session of the Academic English final exam at a California State University campus. There were 8 raters. The data were collected through questionnaires, introspective surveys, and one-on-one interviews with the raters. Each of the interviews lasted from 20 to 30 minutes. The findings of the study are as follows:

Questionnaire Data

Demographic profile of the raters

The following Table shows the demographic profile of each of the 8 raters as collected through a questionnaire:

Rater ID	ESL Teaching Experience	Teaching of Writing Experience	Native Speaker Yes/No	First Language	Self-rating of One's Reading Skill in English*	Experience in Holistic Scoring
Rater 1	2 years	1 year	**No**	**Hungarian**	**Excellent**	1-5 years
Rater 2	10 years	10 years	Yes	English	Excellent	6-10 years
Rater3**	5 years	5 years	Yes	English	Excellent	1-5 years
Rater 4	11 years	12 years	Yes	English	Excellent	6-10 years
Rater 5	7 years	5 years	**No**	**Greek**	**Good**	1-5 years
Rater 6	5 years	3 years	**No**	**Turkish**	**Excellent**	1-5 years
Rater 7	6 years	2 years	Yes	English	Excellent	1-5 years
Rater 8	23 years	3 years	Yes	English	Good	1-5 years

* Excellent/Good/Average/Poor ** Chief rater for the session

Introspective Survey Data

Findings

The raters were given a questionnaire titled 'Holistic Assessment Survey' to fill out before they took part in the interview but while they were half way through scoring the student essays.

Below is the Table containing the total number (frequency) of raters' responses to various questions with regard to their process of reading for assessment? The questions were answered on the scale of 1 to 5 (Strongly agree = 1, Moderately agree = 2, I don't know = 3, Moderately disagree = 4, and Strongly disagree = 5).

Scale (1-5)	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Total
Strongly agree	3	0	1	1	2	1	0	08
Moderately agree	1	0	0	5	3	7	2	18
I don't know	1	2	3	0	0	0	1	07
Moderately disagree	3	4	4	2	2	0	4	17
Strongly disagree	0	2	0	0	1	0	1	04
Total	8	8	8	8	8	8	8	56

Q1: I frequently refer to the scoring criteria while working on a score for an essay ("frequently" is at least once every four essays).
Q2: I have tolerance for grammar errors in the student essay.
Q3: While reading an essay, I specifically look for a thesis statement in it.

Q4: I frequently refer to the other essays I have read whenever possible for maintaining consistency in my awarding of scores.
Q5: I feel that I lose my ability to remain consistent after reading about 15 student essays.
Q6: After reading about 15 essays, I tend to read selectively (i.e., not reading each and every sentence) for awarding a score.
Q7: I frequently refer to the prompt while reading essays.

The introspective survey data indicated that 7 out of 8 raters moderately agreed and one strongly agreed with the statement that after reading about 15 essays, they tended to read selectively for awarding a score (Q6). Note in this regard that many of them (5 out of 8) (62.5%) were concerned that they might lose their ability to remain consistent after a while after a reading of about 15 essays (Q5). The data also suggested that half of the raters (50%) had frequently referred to the scoring criteria (Q1) and most of them (75%) to the other essays that they had previously read while the reading process was in effect (Q4). However, many of them (62.5%) claimed that they did not have to refer to the prompt all the time (Q7). Many of the raters (6 out of 8) (75%) disagreed that they had tolerance for grammar errors (Q2) and half of them (50%) were not sure whether they were consciously looking for thesis statements in the essays they were reading (Q3). The introspective survey data, in general, indicated that raters had frequently referred to either the scoring criteria or the previously evaluated essays for remaining consistent; yet they were afraid that they might lose their ability to remain consistent because of the tiring and the boring nature of the evaluation process. However, they seemed to overcome the tiring process by selectively reading the essays. The findings also indicated that many of them had a good understanding that the essays were not going to be grammatically perfect.

The interview Data

The interview data mostly supported the findings obtained through the questionnaire. The major findings with regard to the human factors involved in holistic evaluation as gathered through the interviews were as follows.

(i) Ability to remember

All raters claimed during the interview that they had to keep remembering how they had graded the previous papers while working on the current ones in order to be consistent. In the words of

Rater 1, “I am weighing an essay against another in my mind.” Rater 7 confirmed this by stating, “You have to keep thinking about ...what did I do before.”

(ii) Inability to avoid subjectivity

The question about whether the holistic assessment process was objective or subjective is quite an important one. It was interesting to note that the answers to this important question varied. Only Rater 7 felt that the tool and the process aimed to be objective. She stated, “if the group is normed well, it is almost an analytical one ... because if everybody can look at the same criteria, then you get it into your head and you should look for the same thing in every essay..."

In contrast to Rater 7, two raters claimed that the tool by its very nature could not be objective at all and the holistic assessment process was purely subjective. Rater 3 wondered whether there was any analytical process in the evaluation process at all. “Analytical process? The rubric is subjective. ...I see in analytical, just count the errors and count the number of words in each paragraph.... there is no analytical in the holistic process.”

Five raters felt that holistic assessment could be a combination of both subjective and objective evaluation. Rater 1 stated that she had to jump back and forth between intuition and her analytical skills: “I have to analyze different parts of the essay for different reasons ... I have to depend on my intuition ... it jumps back and forth, sometimes it is analytical and sometimes it is intuitive.” She complained, however, that she had to get at the score in the end intuitively, and analyzing an essay after she had awarded a score actually affected her judgmental process. Rater 6 agreed with her and stated, “We have a grading hand out so that we know what we are grading for. We are looking for a good positional paper; we are looking for good examples. In that sense, it is an objective process” but “it is on the subjective side ultimately ...More of the subjective and intuitive; we are only human beings ...I think it is in between ... when you feel that you like a paper, I think, then it becomes subjective.” According to Rater 2, it was the criteria that made the instrument and the process objective. In her opinion, her primary focus was to go by the objective criteria as stated in the scoring guidelines “the kind of requirements that are stated in the prompt itself - to summarize and agree or disagree.” She stated that she had to look for a summary and see whether the writer had taken a position with regard to the issue. Yet, she thought that ultimately, the process was subjective

since she had to go by the overall impression the essay had made on her. She concluded her opinion by stating "you do have to not second guess yourself too much. You should read quickly and you know, get an impression, that is what holistic reading is..." There was one rater who did not state her opinion on this issue. The general finding was that in the raters' opinion, though the instrument and the norming process attempted to make the process objective, the process per se was subjective.

(iii) Inability to overcome bias and preferences

As to the question about what they were looking for while scoring the student essays, the raters almost unanimously responded that they were looking for relevant content, proper interpretation of the prompt, clear expression of ideas, consistency of opinion, support for ideas, attention to details, and well-developed paragraphs. Of these, for most of the raters, content came first in their evaluation process though there seemed to be issues related to content versus structure. Their primary concern was that it was possible for an essay that was structurally strong but weak in content to masquerade as a good essay and earn a passing score. Many raters stated that a little bit of weakness in grammar was all right with them. Rater 1 stated that she would fail an essay only if the grammar errors distracted her to the extent that she could not concentrate or focus on the content. For her, if she understood the essay, then she was convinced that the student had produced a passing essay. Rater 2 went even a step further by stating that she would try to instill in her mind at the beginning of every reading session that the student essays were not "going to be perfect grammatically and that most students were not for writing 5 or 6 essays."

Rater 6 was the only one who said that while content was important, "grammar is important too." She unabashedly claimed that she "constantly evaluated the structural parts." She added, "Of course, the use of language matters. If they use a variety of sentence structures, that shows that, they are capable of using the language in different ways."

In summary, it can be said that the raters were looking for different things in the student essays based on what they believed to be important for student writers to show in their writing samples.

(iv) Sympathy for students

Many raters expressed sympathy for the student writers. They seemed to have understood that time as an important variable was working against them. Rater 4 stated her opinion as follows. "In

real class essay, I expect something solid in two weeks, the time they take to think about the topic and write the essay; here, they need to do everything in an hour and a half; standards need to vary between the two." The chief reader brought in another dimension to the whole discussion about time by stating that making students write and produce only the first draft was against the very goal of teaching writing as a process and so the way the students were being tested lacked construct validity.

(v) Other rater variables

Other rater variables, such as mood, background, empathy for student writers, and personal beliefs, also seemed to play a role in affecting one's scoring process. In one nonnative rater's words: "I was not born here and English is my second language, and so I can identify myself with the student writers and their struggles" (Rater 5). The empathetic attitude was not restricted to the nonnative speaker raters alone. Even native speaker raters felt the same, but for another valid reason. In the words of one native speaker rater, his empathy for the student writers prevented him from being objective. He claimed, "The hardest part for me is the objectivity part. And also, I was a student here, so I also realize how much it costs to go to school and the fact that you are failing somebody and they have to take another semester of classes. And so the emotional load is very heavy while scoring these essays" (Rater 8). Rater 1 agreed with him and stated that the hardest thing for her was "not giving a passing score and if they don't pass, the effect it will have on their education. This is a gatekeeping grip." For Rater 5 "scoring process and scoring decision are mood dependent and someone's fate is in my hand."

For Rater 4, an experienced rater, it was easy to get devoured by one's personal beliefs and not to award a student essay what it deserved. She added that it was therefore important that one was able to overcome their personal beliefs and read the student essays with an open mind: "You have to be able to ignore your personal beliefs." Rater 2, another experienced rater, agreed with her and stated that she had her own "personal prejudices" and it was hard to go against these beliefs. Rater 5, a nonnative rater, added that "separating students' opinions from your opinion" was important, "especially when you have strong opinions about the issues discussed in the prompt."

(vi) Propensity to boredom and fatigue due to high number of essays to be evaluated

With regard to the high number of essays one had to grade on the reading day, Rater 4 stated that she would be happy to score only "50 readings a day and not more than that." She added that it was difficult for her to increase the speed "without sacrificing accuracy" since there were too many essays to score. It simply lulled her thinking ability. Fatigue seemed to affect the raters' thinking ability also. Rater 5 reaffirmed this opinion and stated that "grading is boring when there are so many papers. Two days will be better so we can rest a bit." However, Rater 7 had a different kind of opinion on the issue. She said, "there is that factor of getting tired, you get tired, you are fresh at the start of the hour but after a whole hour, at the end of the hour, you are going to be a lot more tired but also you have read a lot so it is easier to speed (sic) and score. I think there are positives and negatives between reading at the beginning and reading at the end." She added that "after taking short breaks and taking snacks or refreshing stuff," she felt that she was somewhat fresh, yet, after a while, somehow she would get tired, and "sometimes, like fallen into rut ... and would like to get it over with."

(vii) Intimidation by the presence of other raters

For 4 of the 8 raters, the presence of other raters and watching how fast they were scoring seemed to have daunted them somehow. In the words of Rater 7, "Not being distracted by the sight of other readers? - When I was first doing it, I would see other people going much faster than I was. I would think, oh, no, I have read only four essays and they have finished a stack of 20. So that is a minor issue; you have to go at your own pace and not pay attention how other people are doing."

(viii) Propensity to halo errors

For Rater 3, the chief reader for the session, there was a cognitive constraint which he called the "Frame of Mind" while scoring the essays. According to him, "everything is relative that you encounter; you encounter an essay that has an accumulation of errors right after you give three essays 5s; with your frame of mind, you become more critical. It's hard to bring in your balanced judgment, especially if you give too many 3s and too many 4s." Note that this rater's opinion is similar to what is called 'halo effect' in the literature. For Rater 4, the cognitive constraint could work within an essay also. In her opinion, if a rater came across a serious error at the beginning of

the essay, it was possible that the rater could suffer from some kind of prejudice about the quality of the paper. It was very important, she claimed, that the rater "should overcome" his/her "prejudice and go for the overall impact of the essay and not be swayed by one error." Rater 7 stated that it was easy not to read an entire essay, especially if the opening paragraphs were not promising. For her, therefore, it was important that one reads the whole essay patiently since "sometimes, it gets much better half way through the essay."

(ix) Propensity to avoid awarding a discrepant score

There were 4 raters who were concerned about awarding a discrepant score while 4 others were not. Rater 4, though she was very experienced, was concerned about the accuracy of her "scoring decision" all the time, and worried that would not "tally with others." Rater 6, a rater with 3 years of experience, in contrast, was not. She would simply follow the standards given to her for scoring the essays and would simply adhere to these.

(x) Inability to decipher certain student handwriting

For at least 3 raters, it was the student handwriting and not the boredom of sitting through the whole reading session that gave them a great deal of headache: "There is an individual factor in handwriting - there is that factor of getting tired; if the essays are typed, it is easier to process, and score fast, I mean it will speed up the scoring process" (Rater 7).

(xi) What was the hardest thing about assigning scores?

To this question, 6 of the 8 raters who took part in the study expressed in one voice that it was the "borderline cases" (Rater 7) that bothered them and gave them a hard time in assigning scores. A borderline essay can be defined as "an essay that has the characteristics of both the failing and passing essays - too many errors but the content reflects the student's ability to critically think on his/her own" (Rater 3). The chief reader (Rater 3) also agreed with the above claim by stating that "it is easy to differentiate a passing essay from a failing essay." Rater 6 agreed with Rater 3 by stating that the hardest part was scoring "3 or 4 papers since it is a pass or fail." For Rater 2 , the hardest thing was differentiating the students' ideas from the content in the prompt especially when they "just write or repeat the ideas in the prompt." For Rater 8, the hardest thing was scoring itself since the scoring criteria were not very specific and were even confusing. In his opinion, there was no

specific guidance as to what he should be looking for: "Are we looking for the ability of a student to be able to present his/her opinion and organize in a cohesive fashion or should it be a combination of the two and should they be equally weighed ..."

(xii) The characteristics of good/ideal raters

The raters were unanimous in defining what makes a good rater. According to them, good/ideal raters are objective and are aware of their own biases; however, they will make sure that their bias does not come in the way of scoring an essay. They also have the ability to look at the big picture and can identify whether or not the writer has the necessary critical skills and convincing logic to support his/her position. They also know that student essays are not going to be grammatical due to the circumstances in which they are written. They are fair in their evaluation and are experienced enough not to hurt students. They are not swayed by the handwriting and they will keep an open mind all the way to the end and then only will decide on the score. They don't get distracted by the sight of other readers. They have an ability to intuitively feel what the writer wants to say. Also, they are familiar with literature on second language acquisition, developmental characteristics of writing an essay, and discourse analysis because they understand that the student writer is not only learning how to write but is also learning the language. Their knowledge of discourse analysis can help them in getting a feel for good organization and cohesion, and also it can provide them with an ability to make out whether the writer has supported his/her position with clear examples. The general opinion of the raters was that they were always conscious of the writer's developmental process. According to them, the ideal rater would also start the evaluation process with the question, how would I answer this prompt? The ideal raters have an ability to think like a student and they have a lot of experience in dealing with different student populations, especially the linguistic and cultural traditions that exist in different ethnic groups. They have a clear grasp of what the standards are and can look at an essay completely and quickly and can make a decision without sacrificing integrity. They can also divorce their readings from their personal idiosyncrasies, their personal preferences for style and format, and their personal prejudices against certain kinds of errors and positions, especially if they have a strong opinion on the issue being discussed. They don't downgrade someone because their position is ridiculous. Finally, ideal raters never get tired; they don't get

distracted by non-crucial things or digressions in the essay; they are able to focus specifically on how the writer presents an argument, and can approach every essay as if it is a new one.

Summary of the major findings of the study:

In summary, in the raters' opinion, the holistic assessment process is neither subjective nor objective but a combination of the two; the nature of the scoring criteria itself is partially responsible for this. The raters also felt that they had to make intuitive judgments in some cases and consciously analyze in other cases. The essays that perplexed the raters also included those student writing samples that proved to be very strong in grammar while lacking in content or vice versa, and those that lie on the borderline, neither qualified to pass nor fulfilled all the failing criteria. The raters stated that they had to remember how they had graded their previous papers in order to remain consistent. The data also suggested that different raters were looking for different things in the essays, and that what one felt as important in a student essay could depend on his/her training and beliefs about writing. It was quite interesting to note that while on the survey raters did not want to admit that they had tolerance for grammar errors, during the interviews, they claimed that they would tolerate errors to the extent that the errors did not distract them from focusing on other aspects of the essay. The raters, furthermore, seemed to be aware that the content of the prompt could affect their rating process and that it was important for them to be aware of their own biases especially with regard to the content of the prompt. The data seemed to indicate that the social and psychological profile of the raters, including their education, upbringing, empathetic nature, understanding of the cost of education, and the stake involved for the students if they did not get a passing score, seemed to affect their scoring process. In the opinion of the raters under study, two of the main reasons that seemed to result in boredom and cause fatigue in the holistic assessment process were student handwriting and long sessions. Some raters also indicated that they did somewhat suffer from anxiety since they were under pressure to avoid discrepancy scores. For some raters, it was difficult to overcome halo errors. The most important finding of the study was that every rater had a good knowledge of what an ideal rater looked like; however, none of them knew how to become one!

Conclusion

The study has shown that the holistic assessment process is not an easy task though it is used in almost all North American universities because of its practicality and efficiency in terms of time. Its widespread use, however, should not be taken as a token of its validity since it has been shown in this paper, through both literature review and empirical findings, that there are many issues that need to be addressed in future research on holistic assessment. What follows are some of the suggestions for improving the validity of the holistic assessment process.

1. One of the important findings of this study indicates that different raters look for different things in their scoring process; therefore, it is important during the norming process that scorers self-explore, with regard to the prompt and the student essays, what they believe in, what they don=t believe in, and what their prejudices and biases are. It is important that raters at the beginning of every holistic assessment session introspect on their prejudices, beliefs, and anxieties so that they can overcome at least some of these in their scoring process.
2. It is important for the chief reader to inform the scorers during the norming process that discrepancy scores are perfectly acceptable so long as there is consistency in each scorer=s scoring process. The raters= tendency to restrict themselves to a certain safe range can thus be avoided.
3. It is equally important for the chief rater to inform the raters during the norming process that they need to work at their own pace and not be deterred or distracted by those who are able to score fast.
4. The scorers also need to be educated that they should read the student writing samples as naturally as possible and make judgments with regard to the score only after they have finished reading the entire paper. It is important that the scores are awarded without much of analysis or reflection since the whole purpose of holistic assessment is to assign a score holistically, and not based on analysis.

5. Efforts should also be made to beguile the boredom that the scorers have to suffer due to long scoring hours. Frequent breaks or even distributing the scoring sessions to two days may, to some extent, can alleviate the boredom and the fatigue of the scorers

In all, the study has clearly shown that there are yet many things that need to be taken care of in order for the process of holistic assessment to become somewhat less intriguing for the raters. However, there is also a serious limitation of this study, in that the findings are based on the data collected from a limited number of instructor-raters who were all teaching the same course on the same campus; as such, it is possible that the findings are specific to this group of instructor-raters. Future research studies should attempt to collect data from a wide range of instructor-raters who teach different writing classes at different levels on different campuses. Such research will test whether what these instructor-raters had been introspecting reflect the same kind of cognitive and affective constraints under which broader populations of instructor-raters carry out the holistic assessment process.

References:

Barritt, L., Stock, P., & F. Clark (1986). Researching practice: Evaluating assessment essays. College *Composition and Communication*, 37: 315-327.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, 18 (1), 65-81.

Erickson, J.D. (2001). *Using keywords and computers to assess student writing*. Dissertation Abstracts International, 61(10), 3964A. (University Microfilms No. DA9988955).

Freedman, S. (1979). How characteristics of students= essays influence teachers= evaluation. *Journal of Educational Psychology*, 71, 328-338.

Freedman, S.W. (1984). The registers of student and professional expository writing: Influences on teachers= responses. In R. Beach and L.S. Bridwell (Eds.), *New directions in composition research*. (pp. 334-347). New York: Guilford Press.

Hamp-Lyons, L. . & Kroll. B. (1996). Issues in ESL writing assessment: An overview. *College ESL*, 6 (1), 56-72.

Huot, B. A. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, 41(2), 201-211.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson and B. A. Huot (Eds.). *Validating holistic scoring for writing assessment*. Cresskill, NJ: Hampton Press.(pp. 206-236).

Legg, S.M. (1998). Reliability and Validity. In W. Wolcott with S.M. Legg. *An overview of writing assessment: Theory, research, and practice*. Urbana, Ill: National Council of Teachers of English. (pp. 124-142).

Pula, J. & Huot, B.A. (1993). A model of background influences on holistic raters. In M. Williamson and B. A. Huot (Eds.) *Validating holistic scoring for writing assessment*. Cresskill, NJ: Hampton Press. (pp. 237-265).

Purves, A. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English*, 26 (1), 108-122.

Shohamy, E., Gordon, C-M.. & Kraemer, R. (1992). The effect of raters= background and training on the reliability of direct writing tests. *Modern Language Journal*, 76(1), 27-33.

Stach, C.L. (1987). *The component parts of general impressions: Predicting holistic scores in college-level essays*. Dissertation Abstracts International, 48(07), 1683A. (University Microfilms No. DA8722706).

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

Williamson, M.M. (1993). An introduction to holistic scoring: The social, historical, and theoretical context for writing assessment. In M. Williamson and B. A. Huot (Eds.). *Validating holistic scoring for writing assessment*. Cresskill, NJ: Hampton Press.(pp. 1-44).

Williamson, M. & Huot, B.A. (Eds.), *(1993). Validating holistic scoring for writing assessment*. Cresskill, NJ: Hampton Press.

Wolcott, W. (1998). Holistic scoring. In Wolcott, W. with S.M. Legg. (1998). *An overview of writing assessment: Theory, research, and practice*. Urbana, Ill: National Council of Teachers of English. (pp. 71-87).

Wolcott, W. with Legg, S.M. (1998). *An overview of writing assessment: Theory, research, and practice*. Urbana, Ill: National Council of Teachers of English.

Zhang, W. (2000). *The rhetorical patterns found in Chinese EFL student writers= examination essays in English and the influence of these patterns on rater response*. Dissertation Abstracts International, 60(09), 3335A. (University Microfilms No. DA9947837)