Available online at www.sciencedirect.com
ScienceDirect
Computers and Composition 25 (2008) 203–223
Computers and Composition
www.elsevier.com/locate/compcom

The Reliability of Computer Software to Score Essays: Innovations in a Humanities Course

A. James Wohlpart [a,*], Chuck Lindsey [b], Craig Rademacher [c]

[a] *Department of Language and Literature, Florida Gulf Coast University, Fort Myers, FL 33965, USA*
[b] *Department of Chemistry and Mathematics, Florida Gulf Coast University, Fort Myers, FL 33965, USA*
[c] *Department of Health, Physical Education, and Recreation, Northern Michigan University, Marquette, MI 49855, USA*

Abstract

In the summer of 2001, Florida Gulf Coast University was awarded a 2-year, $200,000 grant from the National Center for Academic Transformation to redesign a required General Education course entitled Understanding the Visual and Performing Arts. The course redesign project had two main goals: infuse appropriate technology into the course in meaningful ways and reduce the cost of delivering the course. Faculty members in the humanities and arts were adamant that the redesigned course be structured in such a way that it offered a coherent and consistent learning experience for all students and that it maintained the use of essays as an important strategy for learning in the class. The redesign project led to the creation of a wholly online course with all students registered in two large sections. One of the ways in which we continued to incorporate essay writing into the course was to use a computer application, the Intelligent Essay Assessor (IEA) from Pearson Knowledge Technologies, to score two shorter essays. Through detailed assessment, we have demonstrated that the computer software has an inter-rater reliability of 81% as compared to the 54% inter-rater reliability of the holistic scoring by humans. In this essay, we provide general background on the redesign project and a more detailed discussion of the appropriate use and the reliability of the Intelligent Essay Assessor.

Keywords: Essay-scoring software; Inter-rater reliability; Intelligent Essay Assessor; Course redesign; Understanding Visual and Performing Arts

* Corresponding author.
Email addresses: wohlpart@fgcu.edu (A.J. Wohlpart), clindsey@fgcu.edu (C. Lindsey), crademac@nmu.edu (C. Rademacher).

8755-4615/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.compcom.2008.04.001

1. Introduction

The debate over the reliability and usefulness of computer software in scoring essays has been ongoing since the 1960s. In his essay "The Imminence of Grading Essays by Computer—25 Years Later," William Wresch (1993) reviewed the three primary studies of software scoring that had occurred up to that time and concluded that since "no high schools or colleges use computer essay grading. . .there is little interest in using computers in this way" (p. 49). In fact, in his conclusion he stated that while "computer text analysis is still alive and well and making a contribution to our understanding of writing. . .[c]omputer essay grading may actually be less imminent than it was 25 years ago. . ." (p. 57). Wresch's insight into the usefulness of software scoring to assist us in our understanding of how we score essays was quite astute; indeed, this understanding has greatly benefited the development of computer software for scoring essays since the early 1990s.

While computer scoring of essays is not yet as widespread as I sense it may be in the near future, a great deal has changed since Wresch penned (or, more likely, keyboarded) these words. As Lawrence Rudner and Phil Gagne (2001) demonstrated, software has been developed and is being used to score essays with considerable success. In their review of the reliability of three software systems, Project Essay Grade, the Intelligent Essay Assessor (IEA), and E-rater, they concluded, "While recognizing the limitations, perhaps it is time for states and other programs to consider automated scoring services" (p. 5). Jill Burstein and Martin Chodorow (2002) described the advances in computer scoring of essays with a focus on the reliability of the software, a key issue for faculty members as they consider appropriate ways of incorporating such technology into their courses and programs. They conclude, "A goal of current research in automated essay analysis and scoring is to develop applications to ensure that systems maintain a relevant link to what writing experts and test developers believe are critical to the teaching and learning of writing" (p. 497).

In the fall of 2001, Florida Gulf Coast University embarked on a project to redesign a required General Education course entitled Understanding the Visual and Performing Arts. One minor though important aspect of redesigning the course was the idea of using computer software to score two short essays. The faculty members involved in the process struggled with many of the issues that revolve around the use and reliability of essay-grading software. Supported by a $200,000 grant from the Pew Grant Program in Course Redesign through the Center for Academic Transformation (now the National Center for Academic Transformation), the redesign incorporated many of the assignments of the traditional face-to-face course, including four essays, even though the course went fully online. To increase the number of students registered in each section of the course and to reduce costs, the faculty members decided to explore the idea of having two of the four essays that students would write for the course graded by the Intelligent Essay Assessor, a software application developed by Knowledge Analysis Technologies (now Pearson Knowledge Technologies) in Boulder, Colorado. The IEA has its limitations, most notably the length and parameters of the essay assignment and the fact that the software does not provide direct written feedback to students. Because of the skepticism of the humanities and arts faculty involved in the redesign, we focused much of our energy on assessing the reliability of using a computer, rather than humans, to score essays and considered whether this was an appropriate context for using this software. After gathering

and closely analyzing data regarding the validity of the software for four essay prompts and tracking student learning in the course – especially student performance on two longer essays where the students did receive narrative feedback – the faculty found that the software provided a significantly more reliable scoring mechanism than humans and that, in this given and limited situation, the use of the IEA was appropriate.

2. Background and context

Florida Gulf Coast University, the tenth university in the state system in Florida, opened its doors in the fall of 1997 to serve the needs of the southwest Florida region. The mission of the university included an emphasis on the use of technology and the development of courses and programs that could be offered at a distance. Because of our mission, the university created a strong Office of Course and Faculty Development to assist faculty with the use of technology in teaching. In addition, the university attracted faculty who were interested in innovation—in creating new ways of teaching that focused centrally on student learning through the use of university-wide learning outcomes and assessment strategies. We also desired to create a General Education program that would offer coherence across the curriculum at the same time that it offered learning opportunities in specific areas that we deemed essential.

As part of our General Education program, we created a course entitled Understanding the Visual and Performing Arts that focuses on developing the knowledge of the content and contexts of three visual arts (painting, sculpture, and architecture) and three performing arts (music, dance, and theater). Additionally, the course centers on developing the knowledge and skills necessary for engaging the arts as well as a willingness to attend visual and performing arts activities. Students would receive a broad-based understanding of a variety of art forms and develop the skills to analyze and participate in these art forms. The course was created in part to meet the undergraduate university-wide learning outcome of aesthetic sensibility. When the university opened, we were able to meet the demand for this required course with one full-time faculty member and two part-time faculty members, all of whom had a fairly broad training in the arts.

As Florida Gulf Coast University grew, however, we began to experience "course drift." By the 2001–2002 academic year, we were offering 23 sections of the course, all staffed by part-time faculty, many of whom did not have a strong background in both the visual and the performing arts. In addition, we did not have administrative oversight of the course which resulted in a lack of coherence across sections: classes taught by one part-time faculty member with a strong background in the performing arts all but ignored the visual arts; classes taught by another part-time faculty member with a PhD in Art History taught the course as a typical art history survey; classes taught by yet another part-time faculty member with a strong background in the humanities taught the course through poststructuralist theory using Hazard Adams' *Critical Theory Since 1965*. That year, we applied for and received a grant from the Pew Grant Program in Course Redesign, a 2-year, $200,000 grant focused on redesigning courses using technology to increase quality and reduce costs. Significantly, without the redesign of the course, the university would not have been able to continue offering the course because of a lack of appropriate staffing, resources, and classroom space. Because the faculty felt strongly

that this course was one that should continue to be a requirement in the General Education program, they wanted to move forward with the redesign.

3. The redesign project

In order to implement the grant, we assembled a group of faculty members with a broad range of backgrounds—from the various areas of the visual and performing arts and from the humanities. In addition, we had staff from the Office of Course and Faculty Development as key members of the team; they assisted in bringing valid principles of instructional design such as contiguity, reinforcement, and repetition (Gagne, Briggs, & Wagner, 1992) and proven strategies for online curriculum development such as building learning communities and viewing the instructor as a facilitator (Porter, 2004) into the redesign dynamic. The faculty who participated in the redesign project were adamant that we retain two major aspects of the original course: first, that the course would meet the agreed-upon learning outcomes for all students and second, that the course would retain what we saw as two of the most important elements in the course—student visits to theaters and art galleries in the community and the use of essays to demonstrate the ability to apply the material learned in the course to analysis of art works.

The course is now taught online with all students registering in one of two large sections (for a detailed description of the course and the redesign project see Wohlpart, Rademacher, Courcier, Karakas, & Lindsey, 2006). Through the course platform we offered practice tests and module exams on the factual material from the textbook; these multiple-choice tests would assess students' understanding of the content of the course. In addition, however, we retained the use of four essays: two longer critical analysis essays (on student analysis of arts activities that they attend in the community) and two short essays that are a part of the first two module exams (one on the visual arts and one on the performing arts).

To prepare students for writing these four essays, we had them analyze sample essays (some that were strong and some that were weak) in small group discussions that occurred through the course platform. The sample essays were responses to prompts similar to those that the students would see for their critical analysis essays and short essays. As Anne Herrington (1992) has noted, providing students with guidance and feedback as they prepare to write their own essays is an essential element in helping students learn through the process of writing. While such assistance often comes in the form of oral or written feedback from the instructor on initial drafts, Herrington noted that peer feedback and workshops provide another avenue for offering students guidance. In our experience, having students workshop sample essays – some strong, some weak, some "exploded" (these essays include links that explain the various strengths and weaknesses of the sample essays, with comments tied to the scoring rubrics) – allows them to develop their abilities to use the content knowledge they have learned to analyze works of art. In this way, students engage and develop their critical thinking skills and "practice" writing essays even before they write their own essays.

Recognizing that providing students with feedback on essays is an essential part of the learning process, faculty insisted that only the two short essays would be scored by the IEA. While students do not receive written comments on the strengths and weaknesses of their

short essays, feedback was provided in the form of a detailed scoring rubric that allowed them to understand how they could improve their writing. More importantly, the preceptors who oversaw the small group discussions on the samples essays (mostly post-BA students who have degrees in English) graded the two longer critical analysis essays. Anne Herrington (1992) has previously emphasized the importance of the manner of responding to student writing in developing a context for learning. The preceptors who evaluated the longer essays were trained to use the "sandwich approach," opening their narrative comments with a positive statement that discusses the primary strengths of the essay before moving on to a more critical statement that discusses how the essay might have been improved. The narrative comments then conclude with a summative – and positive – statement. As with the short essays, students were also provided with a detailed rubric that provided a rich context for writing these essays, an important element in setting parameters for these types of assignments (Herrington, 1992).

Before deciding to use the Intelligent Essay Assessor to score the two short essays, we researched other essay-grading software on the market. We chose the IEA by Pearson Knowledge Technologies because it operates primarily through content analysis, which fit with the nature of the Understanding the Visual and Performing Arts class—a lower level General Education course, taken primarily by first year students who have only completed the first semester of Composition. The course is not a composition course with an emphasis on writing instruction; rather, it uses writing to further develop student learning of the material. The IEA is most ideally suited for scoring essays that have a narrow range of content (Rudner & Gagne, 2001). The essay assignments in Understanding the Visual and Performing Arts, and especially the short essays that are a part of the module exams, focus on the students' ability to engage the art works through specific terms and concepts outlined in the text—so our primary concern is with the content of what they write, much like that in courses outside of Composition that are content-rich and that are part of a writing-across-the-curriculum program.

In order to score the essays, the Intelligent Essay Assessor uses a "semantic space" (Landauer, Laham, & Foltz, 2000, p. 27) that is created through developing a matrix of relationships among words and passages from 200 holistic scored essays. In addition, an electronic version of the textbook is fed into the computer in order to create a content foundation from the concepts and terms that the students have learned. As Karen Kukich (2000) noted, the Intelligent Essay Assessor measures the quality of writing through Latent Semantic Analysis, which "aims at going beneath the essay's surface vocabulary to quantify its deeper semantic content" (p. 24). As IEA developers Thomas Landauer et al. (2000) explained, "The fundamental idea is that the aggregate of all the contexts in which words appear provides an enormous system of simultaneous equations that determines the similarity of meaning of words and passages to each other" (p. 27). Essentially, through the programming process whereby the software is fed 200+ human scored essays, the IEA develops an "understanding" of the responses to the prompt that is reliant upon correct reference to specific content and materials. In "reading" the scored essays, the IEA learns how to "read" other essays in a holistic fashion, similar to the way in which humans read essays. While appropriate for our work, the Intelligent Essay Assessor does have some limitations. Because it is learning to read essays through analyzing pre-scored essays, the software is most effective with a very narrowly prescribed prompt and with essays that are between 100 and 500 words in length. This fit our needs for the short essays on the exams and might also be used in situations where instructors are assigning short

essays responding to specific prompts in other courses; it would not, however, be applicable in writing situations that are open-ended, such as in our longer essays, or that require extensive feedback, such as in a course that focuses on writing instruction.

At the end of the first year of full implementation of the redesigned course, we used the IEA to score essays for four different prompts: two on the visual arts and two on the performing arts. The prompts that we developed were tied directly to the content found in the textbook that we use for the course, Dennis Sporre's *Reality Through the Arts* (2004). Within each chapter, the text offers students the opportunity to understand the various elements of the arts and assists them in exploring the way in which these elements lead to the creation of "meaning." In the Fall 2002 semester, the short essay question for the first module exam on the visual arts asked students to respond to the following prompt: "Identify the element of color (hue, primary colors, secondary colors, and contrasts) in Henri Matisse's *The Dance*. How does color work to create meaning or experience? What do you think this meaning or experience could be?" Likewise, for the second module exam on the performing arts, students were asked the following: "Identify the formal element of rhythm in 'Spring' from Vivaldi's *The Four Seasons*. How does the rhythm of the selection work to create meaning or experience? What do you think this meaning or experience could be?"

The first part of these three-part questions asks students to apply the factual knowledge that they have gained through studying the text and demonstrated in the multiple-choice portion of the exam to an analysis of a specific piece of art. The second part, however, asks them to apply a higher level of critical thinking skills; students are expected to link their discussion of the specific elements in an artwork to the way in which these elements create meaning. The final part of the questions asks students to apply creative thinking skills, exploring the potential meaning that derives from the first two parts of the question. Thomas Landauer et al. (2000) acknowledged that while the IEA "is typically aimed at factual knowledge" their analysis suggests that it works well even in settings that desire some limited creativity (p. 30). As they noted, because the scoring of the IEA "is based on human judgments of similar essays, the range of performance that the system can measure is unlimited" (pp. 30–31).

The short essay questions for Understanding the Visual and Performing Arts were designed specifically to offer students the opportunity to develop and demonstrate both content knowledge and critical and creative thinking skills. Malcolm Kiniry and Ellen Strenski (1985) outlined eight categories of expository writing in a developmental sequence beginning with less complex writing assignments such as "listing" and "definition" and culminating in more complex assignments such as "analysis" and "argument" (pp. 192–195). They suggest that composition classes should build their assignments towards the more complex and difficult activities. The Composition I class at Florida Gulf Coast University, which is a prerequisite for Understanding the Visual and Performing Arts, works with a similar framework and thus provides a foundation for the writing assignments in the redesigned course, which fall under the category of "analysis" defined by Kiniry and Strenski as "breaking down a text or phenomenon into constituent parts or causes" and, in its most complex manifestations, requiring "an application of some theoretical framework to the object in question" (1985, p. 194). While the short essay questions are rather formulaic, the longer essays are open-ended; students hone their analytical and writing skills on the short essays that have narrower parameters before moving on to the longer essays that invite more exploration.

4. Method

Over the last 5 years, more research has been conducted assessing the inter-rater reliability of computer software when compared to the holistic scoring of humans. One study found that the "automated essay-grading technique ($r = .83$) achieved statistically significant higher inter-rater reliability than human raters ($r = .71$) alone on an overall holistic assessment of writing" (Shermis, Koch, Page, Keith, & Harrington, 2002, p. 16). For our project, faculty members used a holistic scoring process for the initial read of approximately 200 essays for each of the four prompts given in the first year that allowed for the inclusion of all the stakeholders in the redesign project. Our holistic scoring process included a heterogeneous group of faculty – those in the visual arts, the performing arts, and the humanities – because they all had a stake in the redesign process. As Lawrence Rudner and Phil Gagne (2001) noted, "Even with rigorous training, differences in the background, training, and experience of the raters can lead to subtle but important differences in grading." As a result, with our mixed group, we did not expect the typical 70–75% inter-rater reliability that often results from holistic scoring (Rudner and Gagne). Nevertheless, we were hopeful that the computer software would have a higher inter-rater reliability when the computer scores were compared with the human scores.

We developed a rubric based on a four-point scale to score the essays (see Appendix A). While the rubric outlined four different elements of the essays that we were concerned with (Focus, Development, Unity and Coherence, and Grammar and Mechanics), we decided to use a holistic rather than an analytic process because we did not want to separate out and give points or weight to distinct elements in the essay (Moskal, 2000). Indeed, research has generally demonstrated that holistic scoring results in a higher inter-rater reliability than does analytical scoring (Swartz et al., 1999, pp. 502–503). The rubric outlined a "mental model" for scoring essays that would be analyzed and reviewed as the scoring process unfolded over the year; as David Williamson, Isaac Bejar, and Anne Hone (1999) have demonstrated, such a model depends on several factors that lead to an iterative process so that "the advantages of automated scoring may be more fully realized" (p. 163). Within the rubric, a score for a "Middle Range" paper, a score of "3," is dependent on students responding to all three parts of the question in their thesis and in the body of their essays. An "Upper Range" essay, a score of "4," responds to all parts of the question in a highly analytical and creative manner, demonstrating a strong understanding of the content and a strong ability to apply this knowledge through critical and creative thinking skills. A "Lower Range" essay, a score of "2," responds to the question in an inadequate way, not referring to all parts of the question or not developing the response. The lowest score, a "1," is reserved for essays that do not meet the basic criteria for college level writing.

Before we began scoring, the team of faculty members involved in the scoring read 8–10 essays that had been scored by the team leaders, along with several other essays that were not scored. Before each session, the team discussed these sample essays so that we all understood how to use the rubric and score the essays with what we hoped would be some level of consistency. The team spent a great deal of time discussing each essay, comparing it to the scoring criteria on the rubric, and coming to a consensus on the score. As writing teachers can attest, such consensus-style ranking is not the most ideal situation for teaching, as it often hides the differences that enrich our reading experience. Peter Elbow (1993) has argued emphatically

for "*less* ranking and *more* evaluation" in our teaching (p. 188), by which he means a reduction in the assignment of scores or grades and an increase in providing narrative comments and feedback. Clearly, in composition courses or courses that emphasize the teaching of writing, merely providing scores does not assist students in improving their writing. In Understanding the Visual and Performing Arts, which is not a course that teaches writing, the emphasis is on learning the factual information in the textbook and on applying that information to analysis of art works through writing. In this situation, the faculty felt that using a holistic scoring method was appropriate, especially since the students also completed two longer essays for which they did receive narrative comments.

In the scoring sessions themselves, each essay was read by at least two scorers in order to determine the final score of the essay. If an essay received the same score on the first two reads it was considered finished—if not, it went to a third, or sometimes a fourth, read. In our determination of inter-rater reliability, we only considered those readings "reliable" if they provided the same score on the first two reads, thus avoiding the problem of attempting to resolve differences and inflating inter-rater reliability (Cherry & Meyer, 1993, pp. 121–122). Indeed, with a four-point scale (rather than a six-point scale), the general practice is to continue scoring until two readers achieve the same score (with a six-point scale, a one-point differential is often considered acceptable). Once we finished scoring approximately 200 essays for each prompt, the scores were sent to Pearson Knowledge Technologies where they programmed the IEA and then scored all of the student essays submitted for the exam. When the staff at Pearson Knowledge Technologies returned the scores to us, they provided us with a spreadsheet that included each of the readers' scores, the final holistic score (determined by the two readers who were in agreement), and the IEA score. Although our detailed analysis of the reliability of the computer software focused on the prompts given in the first year, we designed the course so that we would have different prompts every semester over a 3-year period.

Using a process similar to that outlined in David Williamson et al. (1999), we reconvened the faculty involved in the scoring process at the end of the first year in order to review the discrepant scores (those where the holistic score and the IEA score differed) and determine whether or not we agreed with our original holistic score or if we would change the score to be in line with the IEA score. We added one external participant, a faculty member versed in assessment that knew about our project but had not been involved in the scoring process during the year; this faculty member was asked to join the group in order to provide an objective perspective as our discussion unfolded during the day. The entire team received copies of all essays that had a score discrepant from the IEA. The essays included the original scores provided by each human reader along with the final holistic score; the team did not, however, know the IEA scores, only that these scores differed from our own. Only the team leader had access to both the holistic scores and the IEA scores; the role of the team leader was to track the discussion and record the final score determined by the team. The scorers went through each essay as a team, discussing the essay and the original scores in relation to the criteria on the rubric; ultimately, a consensus was reached for each essay. The process was complete after all essays with a discrepant score had been reassessed and the final holistic score, whether the same or different from the original holistic score, was determined by the team and recorded by the team leader.

5. Data and analysis

We tracked several elements in our scoring and rescoring process, including the inter-rater reliability of the holistic scoring sessions (Table 1); the breakdown of scores in the holistic scoring sessions (Table 2); a comparison of holistic and IEA scores by category of agreement, which provides the initial inter-rater reliability of the IEA (Table 3); a comparison of holistic and IEA scores by category of agreement after the rescoring session (Table 4); and a comparison of holistic and IEA scores after the rescoring, which provides the final inter-rater reliability of the IEA (Table 5). Following the initial scoring sessions for each prompt and the subsequent programming of and scoring by the IEA, the team discussed the results of the scoring session and compared the results to the scores given by the IEA. These discussions were aimed at further developing and enriching our mental model for the scoring sessions as well as at determining the degree to which we felt that the IEA was a reliable instrument for scoring students' essays.

In our initial holistic scoring sessions, which included essays for two prompts in the Fall 2002 semester and two prompts in the Spring 2003 semester, our team of readers worked together to score the essays using our four-point rubric. These sessions yielded a lower inter-rater reliability than is normal for holistic scoring sessions, primarily because of the heterogeneous nature of the team (Rudner & Gagne, 2001). After reading a total of 803 essays, the team found that 435 were completed after only two reads, yielding a 54% inter-rater reliability among humans for the holistic scoring session. We also found that 317, or 40%, had to be read a third time and 51, or 6%, had to be read a fourth time.

Table 1
Holistic scoring—by agreement

	2 reads	3 reads	4 reads	*N*
Prompt 1	95	84	15	194
Prompt 2	127	70	10	207
Prompt 3	97	90	11	198
Prompt 4	116	73	15	204
Totals	435	317	51	803
Percent (%)	54	39	6	100

While we found that the inter-rater reliability was lower than we would have liked, we agreed that the scoring process was acceptable because of the nature of the scoring (that we considered the score reliable only if the first two scorers provided the same score and that we had a diverse group of readers). An analysis of the process as it unfolded during the course of the year provided feedback that we used to improve our scoring from session to session.

Statistical analysis on these and the following data were carried out using a standard χ^2 test. This test is a statistical measure of the probability that the observed results would be obtained by chance alone, under the assumption that there is no relationship between the variables being reported on in the table. A low probability – p-value – indicates that the observed results are

unlikely to occur by chance, and so provides evidence for a significant relationship between the two variables (Bluman, 2007). Depending on the context, statisticians generally accept a p-value of less than 0.1 (90% level of significance) or less than 0.05 (95% level of significance) as statistical proof of a relationship between the variables. For this essay, we will adopt a 90% level of significance for the purpose of drawing conclusions; following usual practice, the p-values themselves will also be reported so readers can make their own judgments about significance.

For the first prompt, the Fall 2002 Visual Arts Short Essay on the use of color in Matisse's *The Dance*, we scored 194 essays. The human scoring process yielded a very low 49% inter-rater reliability: that is, 95 of the essays were scored the same on the first two reads, while 99 of the essays needed to be read a 3rd or 4th time to reach a consensus. In fact, the inter-rater reliability for the first prompt was significantly lower than the combined reliability of the other 3 prompts, with a p-value of 0.095, meaning that there is less than a 10% probability of the results in Table 1 being obtained by chance alone if the inter-rater reliability for Prompt 1 is the same as for the other three prompts. A debriefing of these results with the faculty members yielded some interesting information. Members of the group suggested that they were surprised at the general weakness of the writing; when they were reminded that this was a lower level course taken almost exclusively by first year students with only one writing course completed, the group recognized that perhaps the expectations with which they scored the essays were too high. One faculty member commented that he "reserved the 4's and 1's, and just focused on giving an even number 2's and 3's"; according to the rubric, however, the primary score that the students should be receiving is a 3—which constitutes a middle range grade, with an equal number of 4's (upper range) and 2's (lower range). The 1's are reserved for essays that do not meet the basic requirements of college level writing and thinking or do not answer the question. The first round of scoring provided us with some interesting insights into the process.

For the second prompt, the Fall 2002 Performing Arts Short Essay on the use of rhythm in Vivaldi's "Spring," we scored 207 essays. With the lessons from the first round of scoring firmly in mind, we increased our own inter-rater reliability in a significant way. The human scoring process yielded a 61% inter-rater reliability: 127 of the essays were scored the same on the first two reads; 80 of the essays needed to be read a 3rd or 4th time to reach a consensus. This reliability was significantly higher than that of the first prompt ($p = 0.039$). The second scoring session occurred 4 weeks after the first and was informed by our understanding of the process from the first session. Because of the detailed debriefing session, we learned better how to use the rubric and score the essays.

For the third prompt, the Spring 2003 semester Visual Arts Short Essay, students were asked a question on the use of form in Paley's *Cross Currents*, a sculpture found on our campus. We scored 198 essays with a 49% inter-rater reliability: 97 of the essays were scored the same on the first two reads; 101 of the essays needed to be read a 3rd or 4th time to reach a consensus. This reliability was not significantly different ($p = 0.670$) from that of the first prompt. Clearly, the inter-rater reliability of the humans dropped once again, this time because of a repeated difficulty that occurred in the students' responses. While the question required students to analyze the use of form in the sculpture, many of them spent a great deal of time discussing the element of color in the sculpture. This very likely occurred for two reasons: first, the sculpture

is vibrant with rich colors, which may have distracted the students, and, second, we may have had the "ghost" of the Fall semester's visual arts question floating around campus, which did ask students to discuss the element of color. As Roger Cherry and Paul Meyer (1993) suggested, inter-rater reliability is not the only source of error in holistic assessment; reliability measures must also include the students and the tests themselves. In this case, while the prompt followed the same format as earlier prompts and came directly from the discussion of meaning found in the text, other factors influenced student performance and the reliability of the holistic session. While some students did link their discussion of color to their discussion of form, we found it difficult to assess what score we would give them based on this type of connection and, as a result, our scores were very uneven.

For the fourth and last prompt, the Spring 2003 Performing Arts Short Essay question, students were asked to discuss the element of dynamics in "Auguries of Spring" from Stravinsky's *The Rite of Spring*. We scored 204 essays, yielding a 57% inter-rater reliability: 116 of the essays were scored the same on the first two reads; 88 of the essays needed to be read a 3rd or 4th time to reach a consensus. Even though this represents a higher percentage than the first prompt, the difference is not at a statistically significant level ($p = 0.271$). Again, our inter-rater reliability for the second scoring session of the semester climbed to an acceptable level probably because of the proximity of the first scoring session and the retention of the lessons learned from the debriefing.

Significantly, for the second and fourth prompts where we had a higher inter-rater reliability, we also found that we consistently scored more essays in the middle range – a 3 – than in the first and third prompts where we had a lower inter-rater reliability. For the second and fourth prompts combined, we gave 249 of the 411 essays a middle range score, about 61%; far fewer, only 103 or about 25% received a 2, a lower range score. For the first and third prompts, which had a lower inter-rater reliability, the team predominantly gave lower range scores. 200 of the 392 essays, or about 51%, were given a 2, a lower range score; far fewer, only 140 or about 36%, were given a 3, a middle range score.

Table 2
Holistic scoring—by scores

	1	2	3	4	*N*
Prompt 1	13	115	58	8	194
Prompt 2	5	45	137	20	207
Prompt 3	10	85	82	21	198
Prompt 4	12	58	112	22	204
Totals	40	303	389	71	803
Percent (%)	5	38	48	9	100

As noted in the discussion above, the second and fourth grading sessions occurred shortly after the first and third sessions, which concluded with lengthy discussions debriefing our experience. The focus of these discussions was on our desire to improve our ability to apply the rubric accurately and consistently. The lessons learned from the first and third sessions carried over into the second and fourth and, interestingly, led to an increase in the percentage of essays scored in the middle range.

While the initial human scoring process yielded a 54% inter-rater reliability, a comparison with the IEA's scores demonstrated a significantly higher inter-rater reliability for the computer scoring. In order to determine the initial inter-rater reliability of the IEA, we compared the final holistic score (the scores determined by humans) with the score given by the IEA. The IEA scored 500 of the 803 essays the same as the holistic scores, a 62% inter-rater reliability. 288 of the essays, or 36%, were scored with a one-category discrepancy (either one higher or one lower), while only 15, or 2%, were scored with a two-category discrepancy.

Table 3
Holistic/IEA scoring comparison—by agreement

	Agree	1 category	2 category	*N*
Prompt 1	105	82	7	194
Prompt 2	146	61	0	207
Prompt 3	118	74	6	198
Prompt 4	131	71	2	204
Totals	500	288	15	803
Percent (%)	62	36	2	100

Again, we also noted a direct correlation between the inter-rater reliability of the holistic scoring sessions and the IEA scores. For the first prompt, the holistic scoring session yielded a 49% inter-rater reliability, as compared to a 54% inter-rater reliability for the IEA (105 of the essays scored by the IEA received the same as the final holistic score; 89 of the essays were scored differently). For the second prompt, the holistic scoring session yielded a 62% inter-rater reliability, as compared to a 71% inter-rater reliability for the IEA (146 of the essays scored by the IEA received the same as the final holistic score; 61 of the essays were scored differently). For the third prompt, the holistic scoring session fell to 49% once again; the IEA scoring process yielded a 60% inter-rater reliability (118 of the essays scored by the IEA received the same as the final holistic score; 80 of the essays were scored differently). For the fourth and last prompt, the holistic scoring session yielded a 57% inter-rater reliability, as compared to a 64% inter-rater reliability for the IEA (131 of the essays scored by the IEA received the same as the final holistic score; 73 of the essays were scored differently). For all but the first prompt, the difference between the holistic scoring and the IEA were significant at our 90% level: the p-values for the four prompts were 0.310, 0.049, 0.034, and 0.099, respectively. For the aggregate of all 803 essays, the IEA inter-rater reliability was significantly higher than the human scoring, with a p-value of 0.001. It is also worth noting that the IEA inter-rater reliability for prompt 1 was significantly different from the aggregate of the other prompts ($p = 0.0075$), suggesting that the experience gained from the first session also positively impacted the training of the IEA scoring program for the subsequent prompts.

Having completed four scoring sessions in the first year of the full implementation of the redesigned course, we reconvened the faculty members as a group in order to analyze the essays that the IEA scored differently than determined in the holistic scoring session by humans. Our review of these essays allowed us to engage the theoretical and the practical foundations of our work and to advance an analysis of the validity of the assessment. As Brian Huot (2002) noted, discussions of validity within assessment of writing has not

advanced as a coherent field, though general parameters have been devised. He asserted from his analysis, borrowing from other researchers, that validity in this type of activity must centrally include the decisions that are made and the actions that are taken as a result of the assessment (Huot, 2002). He concluded that validity should not be perceived as a "pronouncement of approval" but rather as a process that is ongoing that allows critical reflection on the activity and the decisions and actions that result from the activity (p. 51). During the course of our review of these scores, we were able to discuss as a team not only our sense of the accuracy of our own scoring but also the results of those scores. Recognizing that the two short essays, which counted for a total of 10% of the final grade, were only two of four essays written in the course and that the two other essays were longer and were graded, with comments, by humans and counted as 30% of the final grade, we felt confident that the decisions made and actions taken as a result of the assessment were valid.

In order to further our analysis of the validity of the assessment, we looked closely at the 303 essays that were scored differently by the IEA. Of these 303 essays, 288 had a 1-category discrepancy; that is, the IEA scored these essays either 1 higher or 1 lower than the human score. Only 15 essays had a 2-category discrepancy. Of the 288 essays with a 1-category discrepancy, the team decided that 147 of these essays, or about 51%, should receive a different score. In all cases, the scores were changed to be in agreement with the IEA score (that is, none of them were changed to create a 2-category discrepancy). It may seem curious that every essay that was assigned a new score was changed to a score closer to the IEA score, but if we consider that there are only four possible scores, a change in score represents a significant change in the adjudged quality of the essay. It seems unlikely that such a change would move away from the score given by the IEA, which depends on similar human assessment for its training. Of the 15 essays with a 2-category discrepancy, the team decided that 13 of these essays, or about 87%, should receive a different score. In all of these cases, the score was changed to be 1 category closer to the IEA score; none were changed by 2 categories to be in line with the IEA score. As a result of this further critical reflection on the process and the results of the assessment, we determined that the assignment itself, the use of the IEA for scoring, and our use of the scores within the course were valid.

Table 4
Holistic/IEA rescoring—by agreement

	1-Category discrepancy			2-Category discrepancy		
	Human score changed	Human score not changed	*N*	Human score changed to 1 category discrepancy	Human score not changed	*N*
Prompt 1	39	43	82	7	0	7
Prompt 2	41	20	61	0	0	0
Prompt 3	34	40	74	4	2	6
Prompt 4	33	38	71	2	0	2
Totals	147	141	288	13	2	15
Percent (%)	51	49	100	87	13	100

With the additional 147 essays now in agreement with the IEA scores, a total of 647 of the 803 essays were in agreement, yielding a very high 81% inter-rater reliability between the holistic scoring session and the IEA scoring process. Prompts 2 and 4 continued to have the highest inter-rater reliability (90% and 80%, respectively), while prompts 1 and 3 were the lowest (74% and 77%, respectively). The inter-rater reliability after rescoring was not significantly different among prompts 1, 3, and 4. However, the inter-rater reliability for prompt 2 was significantly higher than the others, both individually and in aggregate, with a p-value of less than 0.01 in all cases.

Table 5
Holistic/IEA rescoring comparison—by agreement

	Agree	1 category	2 category	*N*
Prompt 1	144	50	0	194
Prompt 2	187	20	0	207
Prompt 3	152	44	2	198
Prompt 4	164	40	0	204
Totals	647	154	2	803
Percent (%)	81	19	0	100

Nevertheless, the overall inter-rater reliability was sufficient to convince the faculty that in this particular use of the Intelligent Essay Assessor the computerized scoring system was a success. Of the 156 essays that were still discrepant, 122, or about 78%, were scored higher by the IEA than by the human scorers.

The success of the Intelligent Essay Assessor for scoring the short essays in the redesigned Understanding the Visual and Performing Arts course has allowed us to maintain the use of writing in a large enrollment course. Nevertheless, several limitations should be noted concerning our study, similar to those discussed in David Williamson et al. (1999). Because we only evaluated discrepant scores, we do not know if the team would have changed any of the scores that were in agreement. Significantly, in every case where the score was changed, it was brought either into agreement with the IEA (for those with a 1-category discrepancy) or closer to agreement (for those with a 2-category discrepancy). Psychological factors may have also been an issue as the humans knew that the IEA had provided a different score than the human score for the essays they were re-reading. While the group was generally cynical or at least wary of computerized scoring going into the project, after seeing the low inter-rater reliability in the holistic scoring and the high inter-rater reliability from the IEA after each prompt, they had warmed to computerized scoring. This may have influenced their analysis, even though everyone on the team reported a high level of objectivity during the rescoring session. As the day progressed, the team was not apprised of the results; thus they did not know if their decisions were increasing the inter-rater reliability or increasing the level of discrepancy.

A further limitation should be noted, once again, in relation to the software itself. The Intelligent Essay Assessor works best with a very narrowly prescribed prompt and with a short response. The IEA is programmed using scored essays and through this process creates a "semantic space" for "reading" other essays. The software develops an "understanding" of phrases and the relationship between phrases and then scores new essays based on this "under-

standing." As demonstrated by Tim McGee (2006), the IEA can be fooled into giving a high score to what appears to be a nonsensical essay. In McGee's gaming of the IEA, he took a meaningfully and thoughtfully written essay that had been given a high score and reversed the order of the sentences only to find that the new essay, which no longer made sense, still received a high score. Since we did not think that students would so boldly gamble with their grades in this way, we did not believe that we needed to be concerned about students turning in essays written backwards. Significantly, if two students turned in essays that had the same phrases and/or sentences though in a different order, the essays would be flagged for potential plagiarism. What was of greater concern for us was that a student might be able to use certain phrases copied from the course material but not use them in a meaningful way because they lacked an understanding of the material. In our analysis of the essays and the IEA scores, we determined that the IEA consistently gave the highest scores to essays that demonstrated a very high level of understanding of the material and indeed to the essays that were the most thoughtful and creative. The essays that received the lowest scores were those that demonstrated a very low level of understanding of the material and of writing abilities. Nevertheless, McGee's study offers an important caution to the use of computer software in grading essays.

A final concern that has been raised about such studies as ours has to do with the impact of students writing to a machine rather than a human reader. Anne Herrington and Charles Moren (2001) noted that in writing an essay "the writer assumes she will have some impact on the reader's thoughts and feelings" (p. 497). As a result of their own experiences testing computerized software, including the IEA, they concluded that "Writing to a machine... desensitizes us as writers" (p. 497). Clearly, when we teach students about the act of writing, we emphasize the role of audience and purpose in the construction of an essay and we assist students in considering how the words they put on the page (or, more often now, type on their laptop) affect their reader. Yet Herrington and Moran failed to account for the fact that almost all writing at a university occurs in a contrived setting with the primary goal of the activity being the assessment of student learning (of content material, of critical thinking skills, of writing skills), not the determination of whether or not a reader's emotions were affected (how would we grade that?). Even given this shortcoming, though, the issues they have raised concerning the complex relationships between writers and readers – and the changing nature of the writing process itself – on computers, to computers – is an important topic of further research.

6. Conclusion

The integration of writing into courses across the curriculum grew in popularity during the 1970s and 1980s as a way of increasing the emphasis on developing written communication skills. One of the problems that arose with writing-across-the-curriculum programs was the ability of instructors to continue to cover a wide range of course content while adding a new element to their course design (Young, 2003). As Lynette Hirschman, Eric Breck, Marc Light, John Burger, and Lisa Ferro (2000) demonstrated, however, the inclusion of short answer questions, one form of incorporating writing into a course, allows for the development of critical thinking skills that connect the learning of course content to real world applications.

Increasingly, the role of faculty has been "to better develop our [students'] capacity to communicate effectively. This outcome will require working on not only oral and written communication but also critical thinking skills. . ." (Calfee, 2000, p. 35).

Yet with the inclusion of more writing comes more grading, another problem associated with writing-across-the-curriculum programs. Already overburdened, most faculty members do not have the extra time to grade yet another set of essays. Robert Calfee (2000) opened his essay on computerized scoring of essays by noting, "Unfortunately, instructors don't always have sufficient time or resources to effectively grade student compositions or provide feedback on their reading-comprehension skills. This is when automated essay-grading systems. . . can help" (p. 35). To date, writing-across-the-curriculum programs have taken a "cautious approach" to the use of technology, even while writing centers have been more willing to experiment with various computerized systems (Palmquist, 2003, p. 396). Yet the advent of a wide variety of new technologies, including essay-grading software such as the Intelligent Essay Assessor, has provided an opportunity to rethink our curriculum and our strategies for advancing the learning goals that we deem essential for our students (Palmquist).

In content-rich courses such as those in the sciences, social sciences, and even the arts and humanities, the use of computerized software to score essays could allow for a new method – beyond multiple-choice tests – of assessing student learning of material. The use of essays and writing has the added feature of developing critical and creative thinking skills, and the software on the market has demonstrated its ability to reliably score such writing. Indeed, more recent software applications, such as Pearson Knowledge Technologies' Write-To-Learn™, provide students the opportunity to receive feedback on drafts of their essays. With the changes in higher education that we face in our time, the appropriate use of technology can assist us in meeting the challenges that lie ahead: "The greatly expanded diversity of new demands on individuals and institutions will require new and more varied methods of assessing new and more varied aptitudes, competencies, and personal qualities, using new and more varied delivery systems to respond effectively and in timely fashion to new and more varied educational and societal needs" (Messick, 1999, p. 245).

While the decades before the 1990s saw the development and testing of computerized scoring of essays, the time since then has seen the successful design, assessment, and use of such systems. Many states, including Pennsylvania, Indiana, Oregon, and Massachusetts, are experimenting with or already using essay-grading software for standardized tests (Trotter, 2002). Moreover, with the increased expectations for accountability in learning from accrediting bodies such as the Southern Association of Colleges and Schools and the North Central Association of Colleges and Schools, a more concerted effort should be made to create assessment strategies that measure learning as defined by faculty. Samuel Messick (1999) concluded that because "the range of measurement of [student learning] should be expanded and the methods more performance-based. . .a major commitment should be made to computer-based assessment" (p. 252). Rather than allow external pressures such as accrediting agencies or state legislatures to force faculty into untenable situations, they should carefully and thoughtfully implement appropriate technology such as computerized scoring in appropriate settings to meet the learning goals and assessment strategies that they have designed and deem essential.

Appendix A. Hum 2510 Short Essay Scoring Rubric

Scale	Focus/thesis	Development	Unity and coherence	Mechanics
Upper range—strong (4)	• Original and creative thesis that is narrowly focused	• Creative and highly analytical content; excellent observations with strong support	• Essay stays on topic throughout; fully answers all parts of the question	• No spelling, mechanical, or grammatical errors
	• Thesis provides an entrance into the essay that fully answers all parts of the question	• Uses text terminology knowledge-ably and well	• Smooth transitions between topics; strong and coherent flow	
		• Goes beyond description or plot summary	• Meets length requirements	
		• Evidence of enhanced critical awareness, some stretching to issues/concepts beyond the particular work		

Appendix A (*Continued*)

Middle range—good (3)	• Strong thesis that is focused	• Adequate analysis: basic observations with adequate support	• Essay stays on topic but does not always clearly connect the development to the thesis	• Minor spelling, mechanical, or grammatical errors exist but do not impede reading or comprehension
	• Thesis provides an entrance into the essay that fully answers all parts of the question	• Occasional reliance on description or plot summary	• Generally good flow from one topic to another	
		• Good use of text terminology demonstrates understanding of terms	• Meets length requirements	
Lower range—weak (2)	• Thesis exists but is mechanical or expected and not original or creative	• Adequate analysis though may have some obvious or faulty observations; in some places, may lack adequate support; reliance on description or plot summary	• Essay strays from topic on occasion	• Minor spelling, mechanical, or grammatical errors that somewhat impede reading or comprehension

Appendix A (*Continued*)

	• Essay may not answer all parts of the question	• Uses text terminology in a way that demonstrates understanding, though may be limited	• Mechanical flow from one topic to another	
			• Meets length requirements	
Does not meet criteria (1)	• Nonexistent or inadequate thesis	• Little to no analysis: plot summary and/or description only; little to no support for claims	• Essay does not stay on topic	• Spelling, mechanical, or grammatical errors proliferate
	• Essay fails to address the question adequately	• Uses text terminology in a way that fails to demonstrate understanding *or* does not use text terminology	• Awkward format with poor flow between topics	
			• Does not meet length requirements	

A. James Wohlpart is the former Chair of the Division of Humanities and Arts and is currently the Associate Dean in the College of Arts and Sciences at Florida Gulf Coast University. He is a Full Professor of English. He has worked on course redesign through a PEW Grant, and publishes and presents in the area of course redesign and environmental literature.

Chuck Lindsey is currently Associate Professor of Mathematics at Florida Gulf Coast University, where he previously served as director of the general education program during the initial period of this course redesign project. His primary research interests are in the history of mathematics, mathematics education, and computational finance.

Craig Rademacher is a former Instructional Designer at Florida Gulf Coast University. As a creative designer, he currently conducts software instruction for Apple Computer. He is also the owner and principal designer with Instructional Design Services in Fort Myers, Florida. He is currently a faculty member at Northern Michigan University.

References

Bluman, Allan G. (2007). *Elementary statistics: A step-by-step approach* (6th Ed.). New York: McGraw-Hill.

Burstein, Jill, & Chodorow, Martin. (2002). Directions in automated essay analysis. In Robert B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 487–497). Oxford: University of Oxford Press.

Calfee, Robert. (2000). To grade or not to grade. *IEEE Intelligent Systems*, *15*(5), 35–37.

Cherry, Roger P., & Meyer, Paul R. (1993). Reliability issues in holistic assessment. In Michael M. Williamson & Brian A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 109–141). Cresskill, NJ: Hampton Press.

Elbow, Peter. (1993). Ranking, evaluating, and liking: Sorting out three forms of judgment. *College English*, *55*(2), 187–206.

Gagne, Robert Mills, Briggs, Leslie J., & Wagner, Walter W. (1992). *Principles of instructional design*. Fort Worth: Harcourt Brace Jovanovich.

Herrington, Anne J. (1992). Assignment and response: Teaching with writing across the disciplines. In Stephen P. Witte, Neil Nakadate, & Roger D. Cherry (Eds.), *A rhetoric of doing: Essays on written discourse in honor of James L. Kinneavy* (pp. 244–260). Carbondale: Southern Illinois University Press.

Herrington, Anne J., & Moran, Charles. (2001). What happens when machines read our students' writing? *College English*, *63*(4), 480–499.

Hirschman, Lynette, Breck, Eric, Light, Marc, Burger, John D., & Ferro, Lisa. (2000). Automated grading of short-answer questions. *IEEE Intelligent Systems*, *15*(5), 31–35.

Huot, Brian. (2002). *(Re)articulating writing assessment for teaching and learning*. Logan: Utah State University Press.

Kiniry, Malcolm, & Strenski, Ellen. (1985). Sequencing expository writing: A recursive approach. *College Composition and Communication*, *36*(2), 191–202.

Kukich, Karen. (2000). Beyond automated scoring. *IEEE Intelligent Systems*, *15*(5), 22–27.

Landauer, Thomas K., Laham, Darrell, & Foltz, Peter W. (2000). The Intelligent Essay Assessor. *IEEE Intelligent Systems*, *15*(5), 27–31.

McGee, Tim. (2006). Taking a spin on the Intelligent Essay Assessor. In Patricia Freitag Ericsson & Richard H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 79–92). Logan: Utah State University Press.

Messick, Samuel. (1999). Technology and the future of higher education assessment. In Samuel J. Messick (Ed.), *Assessment in higher education: Issues of access, quality, student development, and public policy* (pp. 243–254). Mahweh, NJ: Lawrence Erlbaum Associates.

Moskal, Barbara M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research and Evaluation*, *7*(3) (Retrieved February 9, 2003, from http://pareonline.net/getvn.asp?v=7&n=3)

Palmquist, Mike. (2003). A brief history of computer support for writing centers and writing-across-the-curriculum programs. *Computers and Composition*, *20*, 395–413.

Porter, Lynnette. (2004). *Developing an online curriculum: Technologies and techniques*. Hershey, PA: Information Science Publishing.

Rudner, Lawrence, & Gagne, Phil. (December 2001). An overview of three approaches to scoring written essays by computer. *ERIC Digests* (Retrieved on February 2, 2003, from http://www.ericfacility.net/ericdigests/ed458290.html)

Shermis, Mark P., Koch, Chantal Mees, Page, Ellis B., Keith, Timothy Z., & Harrington, Susanmarie. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement*, *62*(1), 5–18.

Sporre, Dennis J. (2004). *Reality through the arts* (5th Ed.). Upper Saddle River, NJ: Prentice Hall.

Swartz, Carl W., Hooper, Stephen R., Montgomery, James W., Wakely, Melissa B., DeKruif, Renee E. L., Reed, Martha, et al. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement*, *59*(3), 492–506.

Trotter, Andrew (May 29, 2002). States testing computer-scored essays. *Education Week*. Retrieved February 1, 2003 from <http://www.edweek.org/ew/ew_printstory.cfm?slug=38essays.h21>.

Williamson, David M., Bejar, Isaac I., & Hone, Anne S. (1999). "Mental Model" comparison of automated and human scoring. *Journal of Educational Measurement*, *36*(2), 158–184.

Wohlpart, A. James, Rademacher, Craig, Courcier, Lisa, Karakas, Scott, & Lindsey, Chuck. (2006). Online education in the visual and performing arts: Strategies for increasing learning and reducing costs. *Journal of Educators Online*, *3*(1) (Retrieved November 11, 2006, from http://thejeo.com/Archives/Volume3Number1/WohlpartFinal.pdf)

Wresch, William. (1993). The imminence of grading essays by computer—25 years later. *Computers and Composition*, *10*(2), 45–58.

Young, Art. (2003). Writing across and against the curriculum. *College Composition and Communication*, *54*(3), 472–485.