Trent W. Maurer, GEORGIA SOUTHERN UNIVERSITY, tmaurer@georgiasouthern.edu
Jerri J. Kropp, GEORGIA SOUTHERN UNIVERSITY, jkropp@georgiasouthern.edu

The Impact of the Immediate Feedback Assessment Technique on Course Evaluations

ABSTRACT

This project reports the results of two studies that investigated the impact on course evaluations of using partial credit iterative responding (PCIR) with the Immediate Feedback Assessment Technique (IF-AT) forms on summative course assessments. This project also quantifies grade inflation from utilizing different PCIR schemes and documents the percentage of possible partial credit students earned. Study 1 compared evaluations in courses where exams were manipulated. Study 2 compared evaluations in courses where daily reading quizzes were manipulated. Results from Study 1 revealed that multiple course evaluation scores increased 10% in the PCIR condition. Students earned 75% of the partial credit available through PCIR, which resulted in a 10% increase in their exam scores. Results from Study 2 revealed no difference in course evaluations between conditions. Students earned roughly 40% of the partial credit available through PCIR, resulting in a 4 to 8% increase in their quiz scores, depending on the PCIR scheme.

KEYWORDS

Immediate Feedback Assessment Technique (IF-AT), course evaluations, multiple choice exams, quizzes, partial credit iterative responding

This paper reports the results of two studies that investigated the impact on course evaluations of using partial credit iterative responding (PCIR) with the Immediate Feedback Assessment Technique (IF-AT) forms on summative course assessments. This paper also quantifies grade inflation from utilizing different PCIR schemes and documents the percentage of possible partial credit students earned. Although research on the IF-AT forms has been conducted for over a decade, this project is the first to address their relationship to course evaluations and grade inflation.

The Immediate Feedback Assessment Technique, first reported by Epstein, Epstein, and Brosvic (2001), is an alternative to traditional Scantron bubble sheets for recording answers to multiple choice (MC) questions. As described by DiBattista, Gosse, Sinnige-Egger, Candale, and Sargeson (2009, p. 313),

> *The IFAT has an answer-until-correct format and is easily compatible with a variety of grading schemes. The IFAT form has a series of boxes corresponding to the alternatives for a number of MC items. For each item, the*

Copyright © 2015 The International Society for the Scholarship of Teaching and Learning

> *one box associated with the correct alternative has a small star in it, and the other boxes are blank. Boxes are covered by an opaque, waxy coating similar to that found on scratch-off lottery tickets. For each MC item, the student chooses the alternative believed to be correct and scratches the coating off the appropriate box. If a star appears, it confirms that the response was correct, and the student goes on to the next item. However, if the box is blank, the chosen alternative was not correct, and the student can then reconsider the remaining alternatives and continue scratching boxes until the star is uncovered. For each item, the student's final selection will be the correct answer, and if the student completes every item, all of the stars will eventually be revealed.*

The option to select a subsequent answer if the initial answer is incorrect, known as iterative responding (IR), is not a requirement to use the IF-AT forms, as students will still receive feedback that their initial answer is incorrect if they do not reveal the star, but without IR, students who stop after answering incorrectly will not receive feedback about which of the remaining answer choices is correct. (See Dihoff, Brosvic, Epstein, and Cook, 2004; Epstein et al.,2002; and Persky and Pollack, 2008, for visual representations of the forms.) Although IF-AT forms can be used for individual quizzes, exams, and other assessments both formative and summative, their use is not limited to individual settings. IF-AT forms are often used in team-based learning (TBL) settings, where groups of students work together on a joint quiz or activity and complete one IF-AT form as a group (Cotner, Baepler, & Kellerman, 2008; Cotner, Fall, Wick, Walker, & Baepler, 2008b; Lee & Jabot, 2011).

Initial research on the use of IF-AT forms logically focused on their effects on student learning, especially with regard to learning over time (i.e., after receiving feedback). This research demonstrated that compared to students who used Scantrons, students who used the IF-AT forms were significantly more likely to correctly answer similar questions on subsequent assessments (Brosvic, Epstein, Dihoff, & Cook, 2006; Dihoff, Brosvic, & Epstein, 2003; Dihoff et al., 2004; Epstein et al., 2001; Epstein et al., 2002), suggesting that the provision of immediate feedback does indeed boost student learning, particularly if IR is allowed (Brosvic, Epstein, Cook, & Dihoff, 2005; Brosvic et al., 2006). This research also documented that the number of correct first responses on initial assessments typically did not differ between Scantron and IF-AT groups, suggesting that the process of completing an IF-AT form did not itself affect students' likelihood of answering questions correctly on that assessment. That is, the process of receiving feedback about one's initial answers did not affect students' answers to other questions on the same assessment.

As an extension of this wave of research, DiBattista and colleagues explored the use of partial credit with IR (DiBattista, 2005; DiBattista & Gosse, 2006; DiBattista, Mitterer, & Gosse, 2004; DiBattista et al., 2009). Using a scoring system of 100%, 25%, 10%, and 0% for selecting the correct answer on the first, second, third, and fourth attempt, respectively, DiBattista and colleagues documented that students' scores on assessments typically increased by approximately 5 to 6% (e.g., 66% to 71%) as a result of partial credit. Unfortunately, none of these results presented data on the percentage of available possible partial credit students' earned, only the percentage that the partial credit improved their scores. This missing information is critical for two reasons. First, other partial credit

scoring systems are possible (Cotner et al., 2008a; Lee & Jabot, 2011; Persky & Pollack, 2008), and assigning different values to correct answers on the second, third, or fourth attempts would almost certainly result in different increases to scores as a result of partial credit. Failing to standardize the data in terms of how much of the available partial credit students earned makes comparisons across studies more difficult. Second, information about how much partial credit students would likely earn as a result of a partial credit IR system with the IF-AT forms is critical for instructors to know *a priori* when determining the value of second and subsequent answers in order to prevent potentially massive grade inflation. Additionally, within the broader context of authentic assessment of student learning, the percentage of possible partial credit earned gives us valuable new information about student learning that is easily quantitatively expressed. This statistic tells us not just students' scores on assessments, or how many questions students answered correctly, but essentially "how close" students were to the correct answer for the questions they missed on the first attempt. Multiple choice questions typically assess student learning in an all or nothing way, but with information about the percentage of possible partial credit earned, teachers have a much richer and more nuanced descriptive picture of students' mastery of the material.

Additional research explored students' perceptions of the IF-AT forms. This research revealed that students' attitudes towards the IF-AT forms were generally quite positive, particularly compared to Scantron forms (Bowman & Laurent, 2011; Cotner et al., 2008a; Cotner et al., 2008b; Dihoff et al., 2003). Students perceived the IF-AT forms to offer greater clarity in response requirements, a more desirable response format, and greater benefits from testing than Scantrons (Epstein & Brosvic, 2002). Student attitudes were especially positive when the IF-AT forms allowed for iterative responding (Brosvic et al. 2005), regardless of whether partial credit was available or not (DiBattista et al., 2004). Although students thought that MC exams would be fairer with partial credit IR IF-AT forms (DiBattista et al., 2009), over 50% of students preferred the IF-AT forms over Scantrons even when partial credit was not available (DiBattista & Gosse, 2006), which suggests that the effect of partial credit on students' grades cannot be the sole driving factor behind students' preferences for the IF-AT forms.

Teachers do not implement pedagogical changes in a vacuum. Often, pedagogical changes have consequences beyond student learning. One such potential consequence is influences on end-of-course evaluations of teaching. Despite the promising line of research into students' perceptions of IF-AT forms reviewed above, to date, no investigation has explored the relationship between the use of IF-AT forms versus Scantrons and end-of-course evaluations. Given that prior research has documented that students learn more when IF-AT forms are used, that students prefer IF-AT forms (especially with IR and partial credit), and that the prior research has explicitly called for future investigations into the costs and benefits to instructors of using IF-AT forms (DiBattista et al., 2004), this seems a prudent next step in advancing this line of research.

Additionally, the large existing body of research on course evaluations has repeatedly documented a significant relationship between students' expected course grades and student ratings (Franklin, 2001; Ginexi, 2003; Heckert, Latier, Ringwald, & Silvey, 2006), and there is some evidence that this relationship is causal (Maurer, 2006; Salmons, 1993), so the use of IF-AT forms with partial credit IR could significantly affect course evaluations. Further, as Titus (2008) notes, "some researchers have found student ratings to

have unintended negative effects on educational quality through decreasing faculty morale and inducing lowered academic standards and grade inflation (Greenwald and Gillmore 1997a; V. E. Johnson 2003; Ryan, Anderson, and Birchler 1980)" (p. 398). Others have noted that there appears to be an inverse relationship between course difficulty or rigor and course evaluation scores (Addison, Best, & Warrington, 2006). These findings suggest that if the use of IF-AT forms with partial credit IR increases course evaluation scores, this may provide a promising new way to "offset" the effect of increases in rigor on course evaluation scores.

This pilot project arose from a collaboration of two faculty members in the same discipline (Family Science), both of whom were interested in trying the IF-AT forms in their classes as part of their approach to scholarly teaching (McKinney, 2003). After discussing the issues involved, they agreed that this would make for a good Scholarship of Teaching and Learning (SoTL) project that could simultaneously address the gaps in the literature reviewed above, document the percentage of possible partial credit students earn with different IF-AT PCIR systems, and explore the impact on end-of-course evaluations of different IF-AT IR MC assessments. To these ends, this project presents two studies comparing two different types of assessments: a) Study 1 looked at Scantrons vs. high credit PCIR IF-AT, and b) Study 2 at IR IF-AT vs. low credit PCIR IF-AT.

These studies utilize a positivist experimental methodology. We chose this approach for four reasons: a) it was the appropriate method to address the quantitative questions as we had framed them for this investigation, b) it was the method used in all prior investigations of the IF-AT, which would facilitate comparisons to that literature, c) it was the method used in most of the supporting SoTL literature cited in this section, which again would facilitate comparisons, and d) it is widely used in both our own discipline and in our own prior SoTL research.

Based on the prior literature, we hypothesized:

H1: Students using the IF-AT forms with PCIR will earn significantly higher average scores than students using Scantrons or IR IF-AT. However, this difference in scores will be entirely attributable to the impact of PCIR.

H2a: Students using high-credit PCIR IF-AT will give higher course evaluation scores than students using Scantrons.

H2b: Students using low-credit PCIR IF-AT will give higher course evaluation scores than students using IR IF-AT.

GENERAL METHOD

Participants

Participants in both studies were undergraduate students at a rural southeastern doctoral university with an enrollment of 20,000 students.

Materials

Course evaluations were an anonymous, university-mandated common form that included both closed-ended and open-ended items. Nineteen closed-ended items from this form were of interest to this investigation. Six items asked students to compare the course to other courses of similar credit value using a Likert-type rating scale of "1" representing "Much Less" and "5" representing "Much More." An example question was

"How difficult was this course?" One additional course item and 11 instructor items used a Likert-type rating scale with "1" representing "Very Poor" and "5" representing "Very Good." An example item was "The instructor's availability to students was." Because the maximum range for the response items was four, a difference of 0.04 between scores would represent 1%. One final item asked students to report what grade they expected in the course (i.e., A-F).

Procedure

With IRB approval in the "exempt" category, student scores on relevant assessments (exams or quizzes) and course evaluations were collected. Course evaluations were administered during the next-to-last week of class each term. Once the completed anonymous evaluations had been collected by the non-instructor proctor, they were sealed in an envelope, taken to a departmental office, and put in the mail. The instructor received the results of the evaluations approximately two to four weeks after the end of each term. To facilitate comparisons across sections, each course was kept as similar as possible for each condition. No changes were made to assigned readings, lecture content, or review sheets for the duration of the investigation.

STUDY 1

Method

Participants

Participants were 267 undergraduate students enrolled in one of four sections of an introductory Family Development course. Nine students in the Scantron condition and five students in the high PCIR IF-AT condition did not complete the course, yielding a base response rate of 92.80% (N = 116) and 96.49% (N = 137), respectively, exceptionally high course completion rates for this population. Of the students who finished the course, 78.45% (N = 91) in the Scantron condition and 62.04% (N = 85) in the high PCIR IF-AT condition completed anonymous end-of-course evaluations, typical response rates for this population. Demographic information was not collected from participants, but the modal student was a white female who was taking the course as a free elective.

Materials

Students completed three 50-item four-choice MC exams and an end-of-course evaluation. The exams were created by the author, were identical across all four sections, and were composed of approximately 40% factual and conceptual questions and 60% application questions. Exam questions were taken from the assigned readings and from lecture materials.

In the Scantron condition, students indicated their answers to the questions by filling in the corresponding bubble on the Scantron form. In the high PCIR IF-AT condition, students indicated their answers by scratching off the corresponding item on the IF-AT form, using the PCIR scheme devised by Lee and Jabot (2011). If students answered correctly on the first attempt, they earned four points (100%). On the second attempt, they earned two points (50%), and on the third attempt, one point (25%). If they did not answer correctly on the first three attempts, they received zero points for that question. Thus, the maximum score for each exam was 200 points.

Procedure

Student scores on course exams and course evaluations were collected for four consecutive semesters in a single course. The first two semesters used Scantron forms and comprised the Scantron condition. The second two semesters used IF-AT forms and comprised the high PCIR IF-AT condition. Exam scores were posted to the online course management system after exams were scored; exams were not returned. If students wanted to see their exams and what they missed, they were invited to meet with the instructor during office hours.

Results

Hypothesis 1

Student scores across the three exams were compared between the Scantron and high PCIR IF-AT conditions by means of an independent samples t-test. Results revealed that the average percentage exam score for the IF-AT condition (M = 82.19, SD = 6.55) was significantly higher than the Scantron condition (M = 72.36, SD = 9.20), t (199.80) = 9.54, $p < .001$, Cohen's $d = 1.25$. That is, on average, students in the high PCIR IF-AT condition scored almost 10%—or a full letter grade—higher on the exams than students in the Scantron condition.

To determine if this difference was fully or partially attributable to the impact of partial credit available only in the high PCIR IF-AT condition, a follow-up independent samples t-test was conducted. This test compared the average percentage exam score for the Scantron condition (M = 72.36, SD = 9.20) with the average percentage exam score for the high PCIR IF-AT condition *after* any partial credit had been subtracted from the IF-AT exam scores (M = 71.75, SD = 9.50). Results revealed no statistically significant difference between the scores, t (246) = 0.51, *ns*.

To further explore the findings of the first and second t-tests, a paired-samples t-test was conducted comparing the average percentage exam score within the high PCIR IF-AT condition before (M = 71.75, SD = 9.50) and after (M = 82.19, SD = 6.55) the addition of partial credit. Results revealed a statistically significant increase in exam scores after the addition of partial credit, t (133) = 10.44, $p < .001$, Cohen's $d = 1.28$.

One additional descriptive statistic was computed for the high PCIR IF-AT condition, adapted from the formula for normalized learning gain (Hake, 1998): the percentage of possible partial credit earned averaged across all three exams (P_p). Hake's normalized learning gain was an attempt to control for students' pre-existing knowledge in determining "real" learning gains in a course. Consider the case of two students, one who enters a course already knowing 30% of the material to be covered in that course and another who enters the course knowing 0% of the material. If the student with the pre-existing knowledge learned 50% of the material in the course over the duration of the course, that student would finish the course knowing 80% of the course material. If the student without the pre-existing knowledge learned 70% of the course material over the duration of the course, that student would finish the course knowing 70% of the course material. Thus, the student with the pre-existing knowledge finished the course with greater knowledge than the student without the pre-existing knowledge, yet the student without the pre-existing knowledge *actually learned more* in the course, which is obscured by focusing only on end-of-course outcomes. Hake's normalized learning gain controls for students' pre-

Table 1. *Study 1: Course evaluation scores by condition*

	SCANTRON		HIGH-CREDIT PCIR		% IMPROVE-	*F*		PARTIAL
ITEM	*M*	*SD*	*M*	*SD*	MENT	(1, 173)	*P*	η^2
Course difficulty	3.37	.98	2.98	.93	-9.75%	7.34	.007	.04
Instructor preparation	4.74	.68	4.92	.32	4.50%	4.58	.034	.03
Clarity of presentation of material	4.68	.61	4.86	.41	4.50%	5.17	.024	.03
Tests reflected course content	4.38	.88	4.82	.47	11.00%	17.17	.000	.09
Instructor's helpfulness	4.17	1.04	4.62	.62	11.25%	12.29	.001	.07
Instructor	4.37	.84	4.80	.43	10.75%	18.09	.000	.10
Grade	3.03	.76	3.66	.63	15.75%	35.22	.000	.17

existing knowledge by expressing learning gains as a percentage of how much students did not know at the start of the course that they had learned over duration of the course.

In a similar way, the percentage of possible partial credit earned averaged across all three exams (P_p) first calculates how much potential partial credit each student could have earned (students with higher initial scores have fewer incorrect responses and thus fewer chances to earn partial credit), then calculates what percentage of that potential credit they actually did earn. This calculation used the following formula, averaged across all participants:

$$P_p = \frac{S_p - S_r}{(N_t - N_r)\,V} \times 100\%$$

where S_p = score with partial credit, S_r = raw score without partial credit, N_t = total number of questions, N_r = number of questions answer correctly on first attempt, and V = percentage value of a correct answer on a second attempt (e.g., for exams: 1). In this sample, M = 74.90%, SD = 6.99%, range: 57-93%. Thus, on average, students in the high PCIR IF-AT condition earned nearly 75% of the available partial credit.

Hypothesis 2a

A correlation matrix with the 19 course evaluation items was computed and revealed significant correlations between the items. As a result, course evaluation scores across the two conditions were compared with a Multivariate Analysis of Variance [MANOVA] with assessment type as the independent variable and the 19 course evaluation items as dependent variables. A significant multivariate main effect emerged, Wilks' Lamba = .66, $F\,(19, 155) = 4.24$, $p < .001$, partial $\eta^2 = .34$. Follow-up univariate tests revealed significant models for seven course evaluation items. See Table 1, above. Students in the high PCIR IF-AT condition gave significantly better ratings on all seven items than students in

the Scantron condition, supporting Hypothesis 2a. On average, ratings in the high PCIR IF-AT condition were 10% better than ratings in the Scantron condition.

Discussion

Study 1 sought to document the percentage of possible partial credit students earned with a high IF-AT PCIR system and to explore the effect of a high PCIR IF-AT system on end-of-course evaluations. It was hypothesized that students in the high PCIR IF-AT condition would earn significantly higher scores than students in the Scantron condition, but only as a result of the available partial credit. This hypothesis was supported.

Consistent with prior research (Epstein et al., 2001; Epstein et al., 2002), students in the high PCIR IF-AT condition did not score any higher on the exams in their first attempts at questions than students in the Scantron condition, but after adding the partial credit they earned from iterative responses, IF-AT students' scores increased by nearly a full letter grade. Further, subsequent analyses revealed that on average, students in the IF-AT condition earned nearly 75% of the available partial credit. It is important to note that this number should not be interpreted to mean that students selected the correct answer on their second attempt 75% of the time, but only that as a result of second—and third—attempts, students were able to earn 75% of the remaining credit for which they were eligible under this IR scoring system after missing the question on their first attempt.

Hypothesis 2a was also supported. Students in the high PCIR IF-AT condition gave significantly higher ratings on seven course evaluation items than students in the Scantron condition. The average difference between the two conditions was 10%, which is all the more noteworthy given the potential ceiling effect caused by scores that exceeded 4.0 on a 1-5 scale. As expected (DiBattista et al., 2009), students in the high PCIR IF-AT condition perceived the exams to better reflect course content than students in the Scantron condition. Students also perceived the course to be less difficult, despite the fact that it was not objectively any different between the conditions.

Similarly, likely because grades were higher (Franklin, 2001; Ginexi, 2003; Heckert et al., 2006; Maurer, 2006; Salmons, 1993), students in the IF-AT condition rated the instructor more favorably than students in the Scantron condition. Other instructor items that had nothing to do with assessment type (i.e., instructor preparation, clarity of presentation of the material, instructor helpfulness) also showed an increase, suggesting a possible halo effect.

STUDY 2

Method

Participants

Participants were 105 undergraduate students enrolled in one of two sections of an introductory Child Development course. Two students in the IR IF-AT condition and one student in the low PCIR IF-AT condition did not complete the course, yielding a base response rate of 96.67% (N = 58) and 97.78% (N = 44), respectively, representing standard course completion rates for this population. Of the students who finished the course, 82.76% (N = 48) in the IR IF-AT condition and 77.27% (N = 34) in the low PCIR IF-AT condition completed anonymous end-of-course evaluations, slightly above

average response rates for this population. Demographic information was not collected from participants, but the modal student was a white female who was taking the course as a requirement.

Materials

Students completed 20 10-item four-choice MC quizzes and an end-of-course evaluation. The quizzes were adapted by the author from a test bank and were identical across both sections. Quiz questions came from the assigned reading for the day, approximately one half of a textbook chapter.

In the IR IF-AT condition, students indicated their answers by scratching off the corresponding item on the IF-AT form. Students were permitted to keep scratching off items until they revealed the correct answer, but received no partial credit for doing so. In the low PCIR IF-AT condition, students indicated their answers by scratching off the corresponding item on the IF-AT form, using the PCIR scheme devised by DiBattista. If students answered correctly on the first attempt, they earned one point (100%). On the second attempt, they earned 0.25 points (25%), and on the third attempt, 0.10 points (10%). If they did not answer correctly on the first three attempts, they received zero points for that question. Thus, the maximum score for each quiz was 10 points.

Procedure

Student scores on course quizzes and course evaluations were collected for two consecutive semesters in a single course. The first semester used IF-AT forms with IR and comprised the IR IF-AT condition. The second semester used IF-AT forms with low PCIR and comprised the low PCIR IF-AT condition. Quizzes were not returned. If students wanted to see their quizzes, they were invited to meet with the instructor during office hours.

Results

Because both conditions in Study 2 used IF-AT forms with IR, it was possible to compute three sets of scores for each student: a) IR score with no partial credit, b) low PCIR score (using the DiBattista scheme), and c) high PCIR score (using the Lee & Jabot, 2011, scheme). That is, regardless of which scheme was actually used in the course, it was possible to calculate what the students *would have earned* with each potential scoring scheme. Because Study 1 used Scantron forms that do not allow for IR, these comparisons were only possible in Study 2.

Hypothesis 1

Student scores across the 20 quizzes were compared between the IR IF-AT and low PCIR IF-AT conditions by means of an independent samples t-test. Results revealed that the average percentage quiz score for the low PCIR IF-AT condition (M = 64.89, SD = 13.40) was not statistically different from the IR IF-AT condition (M = 66.55, SD = 11.56), t (100) = 0.67, *ns*.

A follow-up independent samples t-test was conducted. This test compared the average percentage quiz score for the IR IF-AT condition (M = 66.55, SD = 11.56) with the average percentage quiz score for the low PCIR IF-AT condition *after* any partial credit had been subtracted from the low PCIR IF-AT quiz scores (M = 60.58, SD = 14.28). Results revealed an unexpected statistically significant difference between the scores,

Table 2.

Study 2: Quiz scores by scoring scheme across condition

	IR IF-AT (N = 58)		LOW-CREDIT PCIR (N = 44)				
SCORE	*M*	*SD*	*M*	*SD*	*T* (100)	*P*	COHEN'S *D*
Actual Score	66.55	11.56	64.89	13.40	0.67	.50	—
IR	66.55	11.56	60.58	14.28	2.33	.02	0.47
Low PCIR	69.49	11.11	64.89	13.40	1.89	.06	—
High PCIR	72.82	10.77	69.83	11.79	1.33	.19	—

Table 3.

Study 2: Differences in quiz scores by scoring scheme within condition

CONDITION	*T*	*P*	COHEN'S *D*
IR IF-AT			
IR vs. Low	-14.52	.000	0.26
IR vs. High	-14.88	.000	0.57
Low vs. High	-14.70	.000	0.31
Low-credit PCIR			
IR vs. Low	-18.22	.000	0.31
IR vs. High	-16.24	.000	0.71
Low vs. High	-11.93	.000	0.40

Note. df *for IR IF-AT is 57*, df *for low-credit PCIR is 43.*

$t\,(100) = 2.33$, $p = .02$, Cohen's $d = 0.47$. That is, students in the IR IF-AT condition answered more questions correctly on the first try than students in the low PCIR condition.

Two additional independent samples t-tests were conducted. The first compared student quiz scores using a low PCIR scheme for both sections. The second compared student quiz scores using a high PCIR scheme for both sections. Neither test was significant. Although students in the IR IF-AT condition answered more questions correctly on the first try than students in the low PCIR condition, the addition of a partial credit scoring scheme—whether low or high—effectively reduced that difference in performance to non-significance. See Table 2.

Paired-samples t-tests were also conducted. The first compared the average percentage quiz score within the low PCIR IF-AT condition before (M = 60.58, SD = 14.28) and after (M = 64.89, SD = 13.40) the addition of partial credit. Results revealed a statistically significant increase in quiz scores after the addition of partial credit, $t\,(43) = -18.22$, $p < .001$, Cohen's $d = 0.31$. Additional paired samples t-tests compared all remaining pairings of scoring schemes for both the IR IF-AT condition and the low PCIR IF-AT condition. All comparisons were statistically significant. See Table 3.

Two additional descriptive statistics were computed for each condition for the percentage of possible extra credit earned: one for the low PCIR scheme and one for the

Table 4

Study 2: Percentage of possible extra credit earned by scoring scheme

	CONDITION					
	IR IF-AT			LOW-CREDIT PCIR		
SCORING SCHEME	M	SD	RANGE	M	SD	RANGE
Low PCIR	37.68%	19.36%	0-89%	45.85%	14.28%	16-78%
High PCIR	39.96%	19.24%	0-83%	47.90%	14.89%	6-81%

high PCIR scheme. In this sample, students earned approximately 40% of the available partial credit. In the low PCIR scheme, scores improved by approximately 3.5%; in the high PCIR scheme, approximately 7.5%. See Table 4.

Hypothesis 2b

Course evaluation scores across the two conditions were compared with a Multivariate Analysis of Variance [MANOVA] with assessment type as the independent variable and the 19 course evaluation items as dependent variables. No significant multivariate main effect emerged, Wilks' Lamba = .68, $F\,(19, 62) = 1.51$, *ns*.

Discussion

Study 2 sought to document the percentage of possible partial credit students earned with a low IF-AT PCIR system and to explore the effect of a low PCIR IF-AT system on end-of-course evaluations. It was hypothesized that students in the low PCIR IF-AT condition would earn significantly higher scores than students in the IR IF-AT condition, but only as a result of the available partial credit. This hypothesis was not supported. Students in the IR IF-AT condition answered more questions correctly on their first attempts than students in the low PCIR IF-AT condition. The effect of partial credit, then, was to increase the scores of students in the low PCIR IF-AT condition to the same level as the scores of students in the IR IF-AT condition. Subsequent analyses revealed that if both sections had used either PCIR scheme, no statistically significant differences in scores would have been detected; partial credit would have effectively "camouflaged" the difference in performance between the classes.

This finding may also have some bearing on Hypothesis 2b, which predicted that students in the low PCIR IF-AT condition would give higher course evaluation scores than students in the IR IF-AT condition. This hypothesis was also not supported. However, this hypothesis was predicated in large part on the assumption that students in the low PCIR IF-AT condition would score higher on quizzes that would raise their course grade which would then translate into higher course evaluations. Because students in the low PCIR IF-AT condition had the same quiz average as students in the IR IF-AT condition, and thus the same grade distribution, these findings suggest that the impact of PCIR on course evaluations may be driven primarily by the contribution of PCIR to grades, rather than other characteristics of the IF-AT forms. Additionally, students in both conditions earned only roughly 40% of the available partial credit under either PCIR scheme, resulting in smaller net influences on scores than in Study 1.

GENERAL DISCUSSION

This pilot project sought to document the percentage of possible partial credit students earned with an IF-AT PCIR system and to explore the effect of an IF-AT PCIR system on end-of-course evaluations. It was hypothesized that students in the IF-AT PCIR conditions would earn significantly higher scores than students in the Scantron and IR IF-AT conditions, but only as a result of the available partial credit (DiBattista, 2005; DiBattista & Gosse, 2006; DiBattista, Mitterer, & Gosse, 2004; DiBattista et al., 2009). This hypothesis was supported for Study 1 for exams, but not for Study 2 with quizzes, largely because of differences between the two conditions in Study 2 in students' raw performance on quizzes. However, this unanticipated result provided unique insight into the potential benefits of using a PCIR IF-AT system. As subsidiary analyses in Study 2 demonstrated, the use of either PCIR scheme (but especially the high PCIR scheme) with both classes would have resulted in essentially the same quiz average for both classes, in essence neutralizing the small but significant difference in raw performance between classes. One possible interpretation of the course evaluation data in Study 2 (Hypothesis 2b) is that students in the low PCIR IF-AT condition might otherwise have rated the instructor lower because of lower grades, but the partial credit they earned which raised their quiz average to that of the IR IF-AT condition students' in turn offset the potential impact on evaluations of lower grades.

If that is the case, it suggests several interesting possibilities for how to use PCIR IF-AT systems to influence course evaluations. As was noted in Study 1, with the support of Hypothesis 2a, students in the high PCIR IF-AT condition gave significantly higher course evaluation scores than students in the Scantron condition. Thus, by changing nothing in the course other than the form on which multiple choice answers are recorded and creating a corresponding partial credit system, instructors may be able to significantly increase their course evaluations, largely due to the impact of partial credit on grade inflation. However, rather than using PCIR IF-AT systems to "game the system" of course evaluations, there may be another practical use that could yield results for students and faculty alike. PCIR IF-AT systems may provide a very meaningful opportunity for instructors to increase rigor in their classes without seeing the corresponding drop in course evaluation scores that typically accompanies lower grades. That is, by offsetting the decline in scores created by more rigorous assessments with the inclusion of PCIR IF-AT forms, the net change to grades could be functionally zero. Rigor increases. Grades are neither inflated nor deflated. Course evaluation scores remain unchanged. This is a "win-win" scenario for students, faculty, and administrators.

One interesting difference between Study 1 and Study 2 was the percentage of possible partial credit earned and the resulting grade inflation. In Study 1, which compared exams, students earned nearly 75% of the available partial credit, resulting in a grade inflation of 10% under the high PCIR scheme. In contrast, in Study 2, which compared daily quizzes, students earned closer to 40% of the available partial credit, resulting in grade inflation between 4 and 8%, depending on the PCIR scheme. In some respects, this is quite surprising, because the exams had 50 questions, but the quizzes had only 10. Yet students were significantly more able to identify the correct answer on subsequent attempts for the longer exams than they were for the shorter quizzes. Further, the exams

covered 1/3 of the material in the course, but the quizzes covered only one half of one chapter of reading. Future research should explore why students seem to be so much better at narrowing down the possible answers to a question on exams than on daily reading quizzes or even if this result can be replicated.

As one reviewer noted, it may be possible that students have different approaches or study strategies for preparing for quizzes than they do for preparing for exams. The data in this investigation cannot directly speak to that possibility, in part because the nature of this project required students to be "blind" to the experimental manipulations, as even making them aware of the nature of the study could have potentially contaminated the results. Informing students that other students in other sections of the course would get partial credit, but that they would not, would very likely have significantly biased any student evaluation scores and rendered any data obtained meaningless. As noted by Felten (2013), although good practice in SoTL preferably involves inquiry into student learning to be conducted in partnership with students, "full partnership may not be practical or appropriate in all SoTL projects" (p. 123). However, future research that more explicitly involves students in the process as co-investigators could qualitatively follow-up on the findings observed here and investigate potential explanations for this seemingly significant difference. It is even possible that just knowing that PCIR is available on the assessments may change student study strategies. For instructors who are considering adopting IF-AT forms with a PCIR scheme, knowing how much of the available partial credit students are likely to earn could be extremely valuable information in determining how to assign partial credit values without creating significant grade inflation or to proportionally offset increases in rigor.

Limitations and Future Directions

This investigation was only a pilot project, and as such, several important limitations need to be noted. First, participants were students in two courses in the same discipline taught by only two instructors at a single university. Although multiple semesters of the courses were used, and the courses were outside of Psychology as recommended by DiBattista et al. (2004), future replication with other courses in other disciplines at other institutions is required before the generalizability of these results can be established. There is insufficient data to know if these patterns of results would generalize to Math or Biology or Chemical Engineering or Art History. Indeed, future replications with other Family Science classes at other institutions is necessary before we can be certain about the results obtained in this investigation. Additionally, because IRB restrictions prohibited the collection of demographic data from our samples, we are unable to fully describe the gender or ethnic makeup of the participants in our study. Modal students in both courses were white females, and based on instructor observation over 90% of the students in both courses were female. Ethnicity is more difficult to assess by observation, but the visible majority in both courses was white, with a sizable minority of African-Americans. Without demographic data, we cannot analyze potential differences by demographics. This is especially important given the gender imbalance in our sample. Future research should explore potential differences by demographics, preferably expanding beyond gender and ethnicity and including other potential variables of interest such as class standing (e.g., sophomores vs. seniors), major, etc.

The second major limitation was that the course evaluation completion rate was lower in the PCIR IF-AT conditions than the Scantron and IR IF-AT conditions. It is possible that the sample for the PCIR IF-AT conditions may have been biased in a way that affected these results. Third, sample sizes were limited, especially in Study 2, which resulted in limited power to detect small effects. Future investigations with larger samples may be unable to uncover smaller effects.

Fourth, because course evaluations were required to be completed anonymously (as is the case at most American colleges and universities), it was impossible to link evaluations with specific students to control directly for course grades or partial credit on IF-AT exams. Future research that can link students' performance with their course evaluations could help disentangle if it is merely the presence of IR/PCIR or the students' specific improvement from partial credit that influenced course evaluations. Fifth, because of logistical limitations, an incomplete experimental design had to be used. As a result, it was not possible to make all possible comparisons between the four assessment methods (Scantron, IR IF-AT, low PCIR IF-AT, high PCIR IF-AT). To more fully explore the nature of the influence of IF-AT forms on course evaluations, future research should explicitly test and compare all of these options.

Trent W. Maurer is an Associate Professor of Child & Family Development in the School of Human Ecology at Georgia Southern University (USA).

Jerri J. Kropp is an Associate Professor of Child & Family Development in the School of Human Ecology at Georgia Southern University (USA).

REFERENCES

Addison, W. E., Best, J., & Warrington, J. D. (2006). Students' perceptions of course difficulty and their ratings of the instructor. *College Student Journal, 40*, 409-416.

Bowman, T. G., & Laurent, T. (2011). Immediate feedback and learning in athletic training education. *Athletic Training Education Journal, 6*, 202-207.

Brosvic, G. M., Epstein, M. L., Cook, M. J., & Dihoff, R. E. (2005). Efficacy of error for the correction of initially incorrect assumptions and of feedback for the affirmation of correct responding: learning in the classroom. *The Psychological Record, 55*, 401-418.

Brosvic, G. M., Epstein, M. L., Dihoff, R. E., & Cook, M. J. (2006). Acquisition and retention of Esperanto: The case for error correction and immediate feedback. *The Psychological Record, 56*, 205-218.

Cotner, S. H., Fall, B. A., Wick, S. M., Walker, J. D., & Baepler, P. M. (2008). Rapid feedback assessment methods: Can we improve engagement and preparation for exams in large-enrollment courses? *Journal of Science Education and Technology*, *17*, 437-443. doi:10.1007/s10956-008-9112-8

Cotner, S., Baepler, P., & Kellerman, A. (2008) Scratch this!: The IF-AT as a technique for stimulating group discussion and exposing misconceptions. *Journal of College Science Teaching*, *37*, 48-53.

DiBattista, D. (2005). The immediate feedback assessment technique: A learner-centered multiple-choice response form. *The Canadian Journal of Higher Education, 35*, 111-131.

DiBattista, D., & Gosse, L. (2006). Test anxiety and the immediate feedback assessment technique. *The Journal of Experimental Education, 74*, 311-327. doi:10.3200/JEXE.74.4.311-328

DiBattista , D., Mitterer, J. O., & Gosse, L. (2004). Acceptance by undergraduates of the immediate feedback assessment technique for multiple-choice testing. *Teaching in Higher Education*, *9*, 17-28. doi:10.1080/1356251032000155803

DiBattista, D., Gosse, L., Sinnige-Egger, J., Candale, B., & Sargeson, K. (2009). Grading scheme, test difficulty, and the immediate feedback assessment technique. *The Journal of Experimental Education*, *77*, 311-338. doi:10.3200/JEXE.77.4.311-338

Dihoff, R. E., Brosvic, G. M., & Epstein, M. L. (2003). The role of feedback during academic testing: The delay retention effect revisited. *The Psychological Record, 53*, 533-548.

Dihoff, R. E., Brosvic, G. M., Epstein, M. L., & Cook, M. J. (2004). Provision of feedback during preparation for academic testing: Learning is enhanced by immediate but not delayed feedback. *The Psychological Record, 54*, 207-231.

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., & Brosvic, G. M. (2002). Immediate feedback assessment techniques promotes learning and corrects inaccurate first responses. *The Psychological Record, 52*, 187-201.

Epstein, M. L., & Brosvic, G. M. (2002). Students prefer the immediate feedback assessment technique. *Psychological Reports, 90*, 1136-1138. doi:10.2466/PR0.90.4.1136-1138

Epstein, M. L., Epstein, B. B., & Brosvic, G. M. (2001). Immediate feedback during academic testing. *Psychological Reports, 88*, 889-894. doi:10.2466/pr0.2001.88.3.889

Felten, P. (2013). Principles of good practice in SoTL. *Teaching & Learning Inquiry, 1* (1), 121-125. doi: 10.2979/teachlearninqu.1.1.121

Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. In K. Lewis (Ed.), *Techniques and strategies for interpreting student evaluation* (pp. 85-100). San Francisco, CA: Jossey-Bass.

Ginexi, E. M. (2003). General psychology course evaluations: Differential survey response by expected grade. *Teaching of Psychology, 30*, 248-251.

Hake, R. R. (1998). Interactive engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*, 64-74.

Heckert, T. M., Latier, A., Ringwald, A., & Silvey, B. (2006). Relation of course, instructor and student characteristics to dimensions of student ratings of teaching effectiveness. *College Student Journal, 40*, 195-204.

Lee, W. T., & Jabot, M. E.. (2011). Incorporating active learning techniques into a genetics class. *Journal of College Science Teaching, 40*, 94-100.

Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology*, *33*, 176-179. doi:10.1207/s15328023top3303_4

McKinney, K. (2003). What is the Scholarship of Teaching and Learning (SoTL) in Higher Education? *Teaching/Learning Matters, 33* (1), 6-7.

Persky, A. M., & Pollack, G. M. (2008). Using answer-until-correct examinations to provide immediate feedback to students in a pharmacokinetics course. *American Journal of Pharmaceutical Education*, *72*, 83. doi:10.5688/aj720483

Salmons, S. D. (1993). The relationship between students' grades and their evaluation of instructor performance. *Applied H.R.M. Research, 4*, 102-114.

Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education, 27*, 397-409.

Titus, J. (2008). Student ratings in a consumerist academy: leveraging pedagogical control and authority. *Sociological Perspectives*, *51*, 397-422.