

# The Exam Autopsy: An Integrated Post-Exam Assessment Model

Leanne R. Havis

Neumann University

Received 14 June 2017; Accepted 3 July 2018

This paper introduces a new integrated post-exam assessment model known as the exam autopsy. Grounded in metacognitive principles of reflective practice, students are provided with three sources of evaluative insight (from self, instructor, and peer) as they seek to analyze the root cause of their exam performance and formulate an action plan for future improvement. The pilot project includes data collected and analyzed over the course of three semesters to chart student performance across two tests using a quasi-experimental design. In Spring 2016 (T1), no metacognitive post-exam intervention was employed. In Fall 2016 (T2), a conventional post-exam self-assessment (or exam wrapper) was used. In Spring 2017 (T3), the exam autopsy model was piloted to provide students with feedback from their instructor and peers in addition to their self-assessment. Statistically significant results from a quantitative analysis of the data suggest that this may be a promising strategy to improve student learning.

## INTRODUCTION

A growing body of literature asserts that encouraging students to use metacognitive practices to monitor, control, and reflect on their own learning can be a vital step in promoting both academic success and the acquisition of transferable skills (Butler, 1997; Downing, 2014; Pintrich, 2000; Zimmerman, 2001). Post-exam self-assessments, or exam wrappers, have been introduced in a variety of formats in higher education courses to stimulate critical thinking among students with a view to improving their individual assessments of their own strengths and weaknesses, as well as their performance on tests. Yet concerns have been raised as to the validity and reliability of self-assessments as true indicators of student ability and progress (Dunn & Mulvenon, 2009; Kruger & Dunning, 1999; Ross, 2006); in other words, students may not know enough to be able to gauge their own ignorance or incompetence, and consequently these may be insufficient as self-monitoring tools.

This paper presents an alternative model for facilitating the reflective post-exam self-monitoring process, one which includes a preliminary self-assessment and a subsequent reflective self-assessment that takes into account both instructor and peer feedback (shared in face-to-face conversations and in writing). Typically, an autopsy is a postmortem examination conducted with the aim of identifying the cause of a person's death or the extent of a pathology or disease. Multiple tests may be carried out, including toxicology, ballistics, or computed tomography, in order to provide insights from different arenas and draw upon various sources of expertise to arrive at an answer. Similarly, this integrated exam autopsy model, vis-à-vis a triangulation of three sources of evaluative insight (self, instructor, and peer) provides a richer basis for students to consider as they analyze the root cause of their poor exam performance, identify whether their current study strategy is indeed working and, if they deem that it is not, to use as a foundation for formulating a new action plan. Moreover, by affording students the opportunity to present their own opinion and judgment of their performance both before and after they are evaluated by others, the instructor is creating additional opportunities for meaningful formative assessment.

The current study presents data collected over the course of three semesters to chart student performance across two tests using a quasi-experimental design. In Spring 2016 (T1), no metacognitive post-exam intervention of any kind was employed. In Fall 2016 (T2), a conventional post-exam self-assessment (or

exam wrapper) was used. In Spring 2017 (T3), the exam autopsy model was piloted to provide students with feedback from their instructor and peers in addition to their self-assessment. Results from a quantitative analysis of the data suggest that this may be a promising strategy to implement in future courses where the instructor is concerned with students' recognition of their own abilities and competencies.

## LITERATURE REVIEW

### Metacognition & Student Self-Awareness

Extensive research demonstrates that the development of metacognitive practices is a vital step in encouraging students to become self-directed or self-regulated learners. Metacognition has been defined as "the process of reflecting on and directing one's own thinking" (National Research Council, 2001, p. 78). It has to do with reflecting on, monitoring, and controlling one's own knowledge and thoughts (Flavell, 1979), and is closely related to self-regulation, which involves the specific skills needed to engage in such reflection, monitoring, and control (Zimmerman & Schunk, 2011). When students are able to assess their own performance effectively, and adapt their approaches or strategies as needed, their learning improves (Delclos & Harrington, 1991). Promoting the improvement of students' metacognitive skills has been shown to result in "not only intellectual habits that are valuable across disciplines (such as planning one's approach to a large project, considering alternatives, and evaluating one's own perspective), but also more flexible and usable discipline-specific knowledge" (Ambrose, Bridges, DiPietro, Lovett, & Norman, 2010, p. 191). When students are able to self-evaluate and self-regulate, they are better positioned to meet specific tasks associated with certain assignments and to pivot (or make different strategic choices) upon recognizing that something is failing to produce the desired results.

Yet one of the most common intellectual challenges faced by students upon entering higher education is managing their own learning (Pascarella & Terenzini, 2005). Warkentin and Bol (1997) found that most students (even upper-division undergraduates) find it difficult to monitor their own efforts. A more recent study determined that, even in cases where students have high expectations for their own performance, they fail to use self-regulation practices consistently (Iwamoto, Hargis, Bordner, & Chandler, 2017). Consequently, it is incumbent upon faculty members to

teach their students how to engage in a variety of processes to monitor and control their own learning (Zimmerman, 2001).

A chief aspect of this self-monitoring process pertains to students' evaluation of their own strengths and weaknesses. Research has shown that "students appear to be especially poor judges of their own knowledge and skills" (Ambrose et al., 2010, p. 195). Moreover, evidence suggests that those students with weaker knowledge and skills are less adept at assessing their own abilities than students with stronger skills (Dunning, 2007). Hacker, Bol, Horgan, and Rakow (2000) asked ninety-nine students to predict their performance in an undergraduate psychology course before and after taking a test and found differing levels of accuracy based on the students' actual performance. Students who scored higher on the test were more accurate in their predictions and postdictions, while students who scored lower "showed gross overconfidence" (Hacker et al., 2000, p. 160) in their performance both before and after the test.

Hacker et al assert that this "lack [of] awareness of their own knowledge deficits" (2000, p. 168) is more significant than any lack of knowledge of course content, insofar as it has serious ramifications for students' abilities to meet their goals. For example, students who believe that they are more prepared for a test than they actually are may spend less time studying, or may use less effective strategies, secure in the flawed knowledge that they are bound to succeed. When they do poorly on a test, they attribute that outcome to "externalizing factors such as a tricky test or unreasonable teacher" (Hacker et al., 2000, p. 168), rather than to any individual practice or trait that they might need to address moving forward.

Dunning, Johnson, Ehrlinger, and Kruger (2003) present additional examples of this lack of self-awareness (Kruger & Dunning, 1999) and attempt to explain the phenomenon with reference to what they call "a double curse" (p. 84). This double curse means that those people who lack the skills to produce the right answers likewise lack the ability to judge whether their own answers (and those of other people) are in fact right. They state that, "In short, incompetence means that people cannot successfully complete the task of metacognition, which, among its many meanings, refers to the ability to evaluate responses as correct or incorrect" (Dunning et al., 2003, p. 85). Such faulty estimates of performance are the result of a "top-down approach" (Dunning et al., 2003, p. 86). People may have a preconceived belief about their skill level in a particular area and then apply those preconceptions to judge how well they are doing on any particular test of that skill. The authors pose the question, "If incompetent individuals do not have the skills necessary to achieve insight into their plight, how can they be expected to achieve accurate self-views?" (Dunning et al., 2003, p. 86).

Downing (2014) explores these ideas within a framework of the language of responsibility, arguing that changing students' inner conversations can lead to key behavior modifications. Downing quotes Nathaniel Branden, who notes that "the object of teaching personal responsibility is to have the student substitute for the question 'Who's to blame?' the question 'What needs to be done?'" (Downing, 2014, p. 50). In his contrasting of the victim and creator mindset, Downing highlights the importance of training students to refrain from making excuses, blaming others, or complaining, and instead to accept ownership of their situation and focus on ways to improve their learning by formulating a plan and taking concrete action. If students are locked into the victim

mindset, they are likely to repeat ineffective behaviors because they are either disinclined or unable to assess their own role in the process accurately. Teaching those necessary metacognitive skills for achieving self-insight, as Dunning et al (2003) posit, becomes all the more critical.

## Self-Assessment & Student Achievement

Bercher (2012) discusses the value of self-assessment and post-exam reflection forms in promoting more accurate self-views among students and fostering the kind of self-regulated learning that could result in more positive outcomes on exams. Seventy-seven students across two semesters in an Anatomy and Physiology class completed the Student Self-Assessment Sheet (SSAS) at the end of each day's lab exercise, expressing their perceived mastery of the content as a percentage. After each of five separate exams, students met with the faculty member and subsequently completed a Post-Exam Reflection Sheet; prompts asked students to determine how much their perceived mastery percentages on the (SSAS) affected their test preparation and to state whether their exam performance matched their expectation of performance. With respect to the former, a majority of the students (87%) reported that the SSAS mastery percentages did impact their exam preparation to some extent. Regarding the latter, 39% of students indicated that they performed as expected on the exam, with 31% indicating that they performed much better than expected and 30% indicating that they performed poorer than expected (Bercher, 2012, p. 29). Bercher concludes that those students who used the information gained from the SSAS became "more aware of their learning strategies and began to feel a sense of control in their ability to choose specific strategies when appropriate" (2012, p. 31). Yet she notes that a small group of students did not adjust their strategies to improve academic performance, "even when faced with certain failure" (Bercher, 2012, p. 31).

It is worth noting that self-assessments may be insufficient as the sole source of data upon which students draw in formulating their judgments and evaluations of effective study strategies. Ross (2006) raises the issue of comparing self-assessments with teacher and peer assessments in the interest of evaluating concurrence rates. He posits that students typically rate themselves higher than their teachers rate them, with some exceptions, and that "agreement of self-assessment with peer judgments is generally higher than self-teacher agreement" (Ross, 2006, p. 3). He suggests that one reason for this may be that students interpret evaluation or assessment criteria differently than their teachers do. According to Ross, self-assessment contributes to student achievement and to improved student behavior, yet its accuracy must be considered and safeguarded (ideally through some triangulation of assessment or evaluation methods which would provide additional perspectives on judgment).

## Peer Assessment & Feedback

Peer assessment between students in higher education has taken a variety of forms (Topping, 1998). Topping (1998), in his review of the literature on a multitude of peer assessment activities, describes the cognitive and metacognitive benefits of "reciprocal same-ability peer assessment, between partners who are equally but differently competent" (p. 254). The process is reflexive, he argues, in that it involves "learning by assessing" (Topping, 1998,

p. 254). For the assessor, spending more time on “cognitively demanding activities” (such as reviewing, summarizing, providing feedback, and filling in gaps) might ultimately “help to consolidate, reinforce, and deepen understanding” (Topping, 1998, p. 254). This requires greater reflection than simply providing a right answer to a question. For the student being assessed, the process provides “swifter feedback in greater quantity” (Topping, 1998, p. 255), but this is only useful when recipients are receptive to feedback. What may facilitate this receptivity is what Topping describes as norm referencing, “enabling a student to locate himself or herself in relation to the performance of peers” (Topping, 1998, p. 255). When students are able to judge their own performance through the lens of their perception of their peers’ performance, they may become better at self-assessing and at identifying the next steps they need to take to improve the quality of their own work.

Gielen and De Wever (2015) contend that peer assessment becomes significantly more effective when peer feedback is scripted. They collected data from 168 first-year undergraduate students who completed a wiki assignment in three cycles of three weeks each. Within each cycle, students were instructed to write a draft version of an abstract for an article, provide peer feedback to (and receive feedback from) another student, and revise the draft version based upon the feedback received. Groups were randomly assigned to a particular condition: the no structure condition, the basic structure condition, or the elaborated structure condition. The no structure condition group received a list of ten predetermined criteria (including problem statement, methodology, results, conclusion, and so on) but was left free to provide feedback as students deemed appropriate. The basic structure condition group received the criteria list and two extra guiding questions: “What do you like about your peers’ work? And “What would you change in your peers’ work?” (Gielen & De Wever, 2015, p. 318). The elaborated structure condition group received a template with specific principles to apply for each criterion on the list. Gielen and De Wever note that “providing structure in the peer feedback template has no influence on the proportion of informative and suggestive elaborations in peer feedback messages between the conditions” (2015, p. 323). They conclude that “adding few guiding questions...significantly increases the elaboration proportion in peer feedback messages, which is beneficial for the peer feedback content quality” (2015, p. 322), but emphasize that the feedback provided should focus on specific criteria rather than on the overall product.

Van den Berg, Admiraal, and Pilot (2006), drawing on earlier work by Topping (1998), present a multiple-case study of seven designs of peer assessment with a view to making a recommendation for an optimal design of peer assessment. They highlight the value of some combination of written and verbal peer feedback in order to maximize the effectiveness of the process. They suggest that “verbal explanation, analysis and suggestions...are necessary elements of the feedback process” (Van den Berg et al., 2006, p. 34), and that two-way feedback should be used (so that the assessor will in turn become the student being assessed). This peer feedback should be exchanged during class time, “because it is difficult to ascertain if students will organize this themselves when out of class” (Van den Berg et al., 2006, p. 35).

Each of these insights about best practices in self- and peer assessment informed the current pilot project, which is described below.

## METHODS

A 200-level criminology course was selected for the current study for two reasons. Firstly, the course is aimed at either second-semester freshmen or first-semester sophomores, although these are not typically the students who enroll in the class (see Table 1, below). Ideally, by targeting students relatively early in their college career, the seeds might be sown for the development of reflective metacognitive skills that may serve them well as they progress toward the completion of their degree. Secondly, the course involves a number of comparatively small-stakes unit exams, rather than simply a midterm and final which are each worth a significant proportion of the overall course grade. Choosing a course with a gap of three or four weeks between each exams was seen as optimal insofar as it affords students enough time to complete the post-exam assessment process (without interfering unduly with the introduction of new course content) but not so much time that the effectiveness of the exam autopsy would have worn off before the next exam was scheduled.

The study followed a quasi-experimental design and took place over the course of three semesters: Spring 2016 (T1), Fall 2016 (T2), and Spring 2017 (T3). T1 essentially functioned as the control group; no interventions or post-exam assessments of any type were implemented between the first and second tests. In accordance with institutional review board protocol, students were notified that their exam scores would be included in the current research project and were given informed consent forms to sign; declining to participate would mean that the students’ scores for the first and second exams would be excluded from the data analysis process. No student declined to participate (n=29). Students’ scores for each exam were noted, along with the mean scores and standard deviation, as were the percentage changes. These are presented below.

During T2, a traditional exam wrapper or post-exam assessment model was introduced after students received their grades for the first exam. Since the exam was administered online through the course management system, students could view their results as soon as the fill-in-the-blank questions were reviewed and scored. At the start of the following class period, students were told that this post-exam self-assessment would be taking place and that the objective of the assignment was for them to think critically about their study strategies and to identify opportunities for improvement. Students were given the informed consent forms to sign at this time; they were told that if they declined to participate in the research project, they would still need to complete the post-exam self-assessment assignment, but that their scores for both the first and second exams, and any data associated with the work they handed in, would be excluded from the study. No student declined to participate (n=22). Class time was set aside for students to review correct and incorrect answers on the test and to address the following questions in writing:

- How did your actual grade on this exam compare with the grade you expected? How do you explain the difference, if there is any?
- How do you feel about your exam grade? Are you surprised, pleased, relieved, disappointed, or what?
- How many hours did you spend preparing for this exam? Was this enough time to get the grade you wanted, or should you have spent more time preparing?

- How did you spend your time preparing for the exam? (For instance, did you summarize your notes? Did you make and use flash cards? Did you test yourself in some way? Did you study with classmates?)
- Examine the items on which you lost points and look for patterns. Did you misread the questions? Were you careless? Did you run out of time? Did you think that you wouldn't need to study as much as you would for an in-class exam since you could use your notes?
- Set a goal to get a certain percentage correct in the next exam. What study strategies and schedule will enable you to earn that score?

Students turned their post-exam self-assessment responses in for the faculty member to review, and approximately fifteen minutes were spent at the start of the following class period discussing those areas of concern that the faculty member identified as common across a majority of the students. The faculty member also introduced some information about the effectiveness (or ineffectiveness) of particular study skills, and encouraged students to seek out a peer tutor in the Academic Resource Center if they felt that their concerns about exam anxiety or note-taking required more extensive remediation than the faculty member could provide. Students took the second exam four weeks after the date of the first exam. Their scores for both the first and second exam were again noted, along with the mean scores and standard deviation, as were the percentage changes. These too are presented below. It should be noted that students received credit for completing the post-exam self-assessment, but not a merit-based grade.

During T3, the exam autopsy model was piloted. Students again took the first exam through the course management system, and were able to view their results as soon as the fill-in-the-blank questions were reviewed and scored. At the start of the following class period, they were informed about the exam autopsy process. The faculty member explained the steps involved and encouraged students to think deeply and honestly about their study strategies and possible opportunities for improvement. The faculty member distributed and collected informed consent forms at this time, and students were notified that if they declined to participate in the research project, they would still need to complete the exam autopsy process as a course requirement; however, their scores for both the first and second exams, as well as any data associated with the worksheets they handed in, would be excluded from the study. No student declined to participate ( $n=23$ ). It should be noted that students seemed to appreciate and find humor in the fact that the process was called an "exam autopsy." The idea that they would be afforded the opportunity to dissect and investigate the root causes of their exam performance from an objective, almost detached position (not unlike that of a detective or coroner, as they described it), was highly appealing. For that reason, the model retains its original name.

The faculty member set aside class time once again for students to review correct and incorrect answers on the test and to address the aforementioned questions in writing. This time, however, the preliminary self-assessment responses were not immediately turned in for the faculty member's review. Instead, students were paired up with a partner who served as a peer evaluator. Partners were assigned randomly. Following a brief lecture-based session about the do's and don'ts of providing feed-

back (i.e., begin by saying something positive about the efforts of the person being evaluated, limit suggestions about areas of improvement to three to avoid overwhelming the person being evaluated, and ensure that all comments are constructive rather than derogatory), pairs of students began working collaboratively to review each other's answers. In their capacity as peer evaluators, students were instructed to write down comments about each of their partner's answers. Specifically, they were asked to consider whether their partner's assessment was valid, whether their partner's goals were realistic, and whether they was anything else they felt their partner should consider. The faculty member provided students with the following follow-up questions, which they were asked to respond to in writing and then to share with their partner out loud, expanding on anything that may have been unclear or vague:

- Do you agree with your partner's assessment of how and why s/he earned a different grade than expected? Why or why not?
- Any and all feelings your partner may express about his/her exam grade are valid. What words of wisdom or comfort could you share in light of how s/he feels?
- What is your opinion of the time your partner spent studying for this test?
- What is your opinion of the methods your partner used in studying for this test?
- What is your opinion of your partner's assessment of the questions s/he got wrong? Do you have another interpretation of or explanation for what might have happened?
- What do you think of the goals that your partner has set for him/herself? Are they realistic? What are two additional ideas you could suggest to help him/her achieve those goals?

The peer conferencing session took approximately twenty minutes of class time. Once it was concluded, students turned in their preliminary self-assessment and peer feedback worksheets (attached to one another). The faculty member then reviewed each of these outside of class and provided feedback on a third worksheet, addressing the following prompts:

- Do I agree with your assessment of why you got a different grade than expected? Why or why not?
- Any and all feelings you may express about your exam grade are valid. What words of wisdom or comfort could I share in light of how you feel?
- What is my opinion of the time you spent studying for this test?
- What is my opinion of the methods you used in studying for this test?
- What is my opinion of your assessment of the questions you got wrong? Do I have another interpretation of or explanation for what might have happened?
- What do I think of the goals that you have set for yourself? Are they realistic? What are two additional ideas I could suggest to help you achieve those goals?

Rather than providing these written comments to students at the start of the next class period, the faculty member instructed students to sign up for a five- to ten-minute individual face-to-face meeting sometime during the following week to discuss their worksheets. The purpose of the face-to-face meeting was to elaborate on, and provide clarification for, any aspects of the faculty feedback that was vague, ambiguous, or confusing, and students were then given their three worksheets (all attached)

to take with them in preparation for the final step of the exam autopsy process. For that final step, the reflective self-assessment, students were given the following instructions:

Think about your original answers to the self-assessment questions, as well as the feedback that you received from your partner and from me. In a brief paragraph, write down what, if anything, has changed in terms of how you prepared for the first test and how you plan to prepare for the next test. Be concrete and specific in describing at least three strategies that you plan to use to study for (or take) the next test. Why do you think those strategies are the most promising for you? What can I do to help support your learning and your preparation for the next exam?

In all, the exam autopsy process took ten days from start to finish. During that time, the faculty member continued to introduce course content both in and out of class (using the course management system). The second exam was administered four weeks after the date of the first exam. Students' scores for both the first and second exam were noted, along with the mean scores and standard deviation, as were the percentage changes. These are presented below. It should be noted that, as in T2, students received credit for completing the exam autopsy process, but not a merit-based grade for the worksheets themselves.

Some demographic data about the sample for each of the time periods under study may be helpful for visualization purposes, and also to posit overall group equivalence (see Table 1). In T1 and T3, the samples were comprised of more male than female students; in T2, female students made up the majority. All three samples included a predominantly Caucasian student body, albeit to varying degrees. Students who self-identified as Hispanic accounted for less than 10% of each sample. There was some variability in students' class level. In T1 and T2, seniors made up an overwhelming majority of the class; in T3, class levels were more evenly distributed, with sophomores, juniors, and seniors all accounting for approximately the same percentages.

## RESULTS

In T1 ( $n=29$ ), the mean score students earned on the first exam was 74.21 (with a standard deviation of 17.50) and the mean score earned on the second exam was 60.28 (with a standard deviation of 26.54). Once percentage changes were calculated for the entire class, data showed student scores dropped by an average of 14.72% (with a standard deviation of 41.42). Part of the reason for the tremendous variability is that two students out of twenty-nine failed to take the second exam, and the zero that was recorded as their exam grade was included in the calculations. When those students' scores for both exams were excluded ( $n=27$ ), the mean score earned on the first exam changed to 73.59 (with a standard deviation of 17.33) and the mean score earned on the second exam was 64.74 (with a standard deviation of 21.41). Percentage changes were recalculated for the entire class and data showed that student scores still dropped, but only by an average of 8.40% (with a standard deviation of 35.31).

In T2 ( $n=22$ ), students earned a mean score on the first exam of 77.23 (with a standard deviation of 10.67) and a mean score on the second exam of 60.77 (with a standard deviation of 14.28). Once percentage changes were calculated for the entire class, data showed student scores dropped by an average of 20.73% (with a standard deviation of 19.37).

**Table 1. Demographic Data for Students in T1, T2, and T3**

|                       | T1 | %      | T2 | %      | T3 | %      |
|-----------------------|----|--------|----|--------|----|--------|
|                       |    | (n=29) |    | (n=22) |    | (n=23) |
| <b>Sex/gender</b>     |    |        |    |        |    |        |
| Male                  | 17 | 58.6   | 10 | 45.5   | 14 | 60.9   |
| Female                | 12 | 41.4   | 12 | 54.5   | 9  | 39.1   |
| Total                 | 29 | 100%   | 22 | 100%   | 23 | 100%   |
| <b>Race/ethnicity</b> |    |        |    |        |    |        |
| African American      | 12 | 41.4   | 6  | 27.3   | 9  | 39.1   |
| Caucasian             | 17 | 58.6   | 16 | 72.7   | 14 | 60.9   |
| Total                 | 29 | 100%   | 22 | 100%   | 23 | 100%   |
| Hispanic              | 1  | 3.4    | 1  | 4.5    | 2  | 8.7    |
| Non-Hispanic          | 28 | 96.6   | 21 | 95.5   | 21 | 91.3   |
| Total                 | 29 | 100%   | 22 | 100%   | 23 | 100%   |
| <b>Class level</b>    |    |        |    |        |    |        |
| Freshman              | 3  | 10.3   | 1  | 4.5    | 2  | 8.7    |
| Sophomore             | 4  | 13.8   | 0  | 0      | 8  | 34.8   |
| Junior                | 3  | 10.3   | 4  | 18.2   | 6  | 26.1   |
| Senior                | 19 | 65.5   | 17 | 77.3   | 7  | 30.4   |
| Total                 | 29 | 100%   | 22 | 100%   | 23 | 100%   |

In T3 ( $n=23$ ), the mean score students earned on the first exam was 74.87 (with a standard deviation of 11.15) and the mean score earned on the second exam was 80.17 (with a standard deviation of 5.87). Once percentage changes were calculated for the entire class, data showed student scores improved by an average of 8.81% (with a standard deviation of 13.94). This was the only group out of the three populations being studied that showed an average increase, rather than decrease, between the first and second tests (see Table 2).

**Table 2. Mean Scores Across Exams in T1, T2, and T3**

|                  | Exam 1 Mean | Exam 1 Std. Dev. | Exam 2 Mean | Exam 2 Std. Dev. | Mean % Change from Ex 1 to Ex 2 | Std. Dev. of Change from Ex 1 to Ex 2 |
|------------------|-------------|------------------|-------------|------------------|---------------------------------|---------------------------------------|
| <b>T1 (n=29)</b> | 74.21       | 17.50            | 60.28       | 26.54            | -14.72%                         | 41.42                                 |
| <b>T1 (n=27)</b> | 73.59       | 17.33            | 64.74       | 21.41            | -8.40%                          | 35.31                                 |
| <b>T2</b>        | 77.23       | 10.67            | 60.77       | 14.28            | -20.73%                         | 19.37                                 |
| <b>T3</b>        | 74.87       | 11.15            | 80.17       | 5.87             | 8.81%                           | 13.94                                 |

A baseline comparison of first exam scores across all three semesters was conducted using a single-factor analysis of variance (ANOVA) in order to determine whether statistically significant differences existed prior to the introduction of any post-exam intervention (see Table 3). The average grades earned by students on each of the first exams did not differ in a statistically significant way; for all intents and purposes, they were fairly equivalent.

**Table 3.** Single Factor ANOVA – Differences in First Exam Grades

| SUMMARY             |          |      |          |          |          |          |
|---------------------|----------|------|----------|----------|----------|----------|
| Groups              | Count    | Sum  | Average  | Variance |          |          |
| T1 (Spring 2016)    | 29       | 2152 | 74.2069  | 306.0985 |          |          |
| T2 (Fall 2016)      | 22       | 1699 | 77.22727 | 113.803  |          |          |
| T3 (Spring 2017)    | 23       | 1722 | 74.86957 | 124.3004 |          |          |
| ANOVA               |          |      |          |          |          |          |
| Source of Variation | SS       | df   | MS       | F        | P-value  | F crit   |
| Between Groups      | 120.6204 | 2    | 60.3102  | 0.312665 | 0.732497 | 3.125764 |
| Within Groups       | 13695.23 | 71   | 192.8906 |          |          |          |
| Total               | 13815.85 | 73   |          |          |          |          |

A single-factor ANOVA (with a significance level of 0.05) was subsequently performed on the changes in students' exam scores for all three populations, and found that there was a statistically significant difference depending upon the type of post-exam assessment utilized (see Table 4), even when the number of students included in the Spring 2016 class (T1) was adjusted (from  $n=29$  to  $n=27$ ) to account for the two individuals who did not take the second exam (see Table 5).

A post-hoc analysis was conducted using the Tukey procedure to test all pairwise comparisons. The aforementioned ANOVA revealed that there were statistically significant differences across the three groups but did not clearly indicate where those differences lay. While the HSD statistic for T1 and T2 was not greater than the critical value of 2.83, those for the two pairings involving T3 (namely, T1 and T3, and T2 and T3, respectively) were greater (3.386 and 5.51, respectively). This suggests that the exam autopsy process did result in statistically significant differences in student performance on the second exam.

## DISCUSSION

Assuming that the three groups of students across the three semesters under study were equivalent, and that the only difference that could have accounted for grade changes between the

**Table 4.** Single Factor ANOVA when Spring 2016  $n=29$ :

| SUMMARY                             |         |              |          |          |          |  |
|-------------------------------------|---------|--------------|----------|----------|----------|--|
| Groups                              | Count   | Grade Change | Variance |          |          |  |
| T1 - No post-exam assessment        | 29      | -14.71655172 | 1715.426 |          |          |  |
| T2 - Post-exam self-assessment only | 22      | -20.72590909 | 375.0304 |          |          |  |
| T3 - Exam autopsy                   | 23      | 8.814782609  | 194.2786 |          |          |  |
| ANOVA                               |         |              |          |          |          |  |
| Source of Variation                 | SS      | MS           | F        | P-value  | F crit   |  |
| Between Groups                      | 11269.4 | 5634.699653  | 6.647597 | 0.002258 | 3.125764 |  |
| Within Groups                       | 60181.7 | 847.6295206  |          |          |          |  |
| Total                               | 71451.1 |              |          |          |          |  |

first and second exams was the type of post-exam intervention introduced (if at all), it would seem that the exam autopsy procedure implemented in T3 (Spring 2017) is a useful and significant tool that faculty members can use to promote self-regulated learning and meta-cognitive reflection among their students. The group of students that employed the exam autopsy approach in T3 was the only one of the three under study to see an overall improvement in test scores between the first and

second exam.

There are two main limitations with the current pilot project. The first has to do with internal validity. It is impossible to determine with any certainty that the results calculated using the analysis of variance are indeed wholly attributable to the structure and format of the exam autopsy model itself and not to the particular cohort in any given semester or to testing effects. It is possible that students in T3 were simply more motivated to improve (i.e., that they were "better students"), or that being subjected to such a rigorous post-exam evaluation process impelled them to invest more time and effort into studying for the second exam. Although the comparison of first exam scores presented above would suggest that the groups were equivalent at the outset, no attempt was made to measure student motivation levels or grasp of study skills. In other words, it is possible that students in T3 were naturally more self-aware and/or more driven to invest more effort into studying for a second exam (and also knew how to study more effectively) once they recognized that their study strategy needed to be modified.

The second limitation has to do with the quantitative nature of the research. Soliciting student comments about how effective they felt the exam autopsy process was and what changes, if any, they had made to their study strategies, either immediately following the second exam or at the end of the semester, could have been highly illuminating.

Nonetheless, given that this exam autopsy model did in fact produce statistically significant changes in students' exam performance, three future research opportunities present themselves. Firstly, the pilot project utilized what Topping (1998) terms "reciprocal same-ability peer assessment." That is, students were random-

**Table 5.** Single Factor ANOVA when Spring 2016 n=27:

| SUMMARY                             |          |         |              |          |         |          |
|-------------------------------------|----------|---------|--------------|----------|---------|----------|
| Groups                              | Count    | Sum     | Grade Change | Variance |         |          |
| T1 - No post-exam assessment        | 27       | -226.78 | -8.399259259 | 1246.457 |         |          |
| T2 - Post-exam self-assessment only | 22       | -455.97 | -20.72590909 | 375.0304 |         |          |
| T3 - Exam autopsy                   | 23       | 202.74  | 8.814782609  | 194.2786 |         |          |
| ANOVA                               |          |         |              |          |         |          |
| Source of Variation                 | SS       | df      | MS           | F        | P-value | F crit   |
| Between Groups                      | 9942.152 | 2       | 4971.076237  | 7.69799  | 0.00096 | 3.129644 |
| Within Groups                       | 44557.64 | 69      | 645.762926   |          |         |          |
| Total                               | 54499.79 | 71      |              |          |         |          |

ly paired up with one another with no consideration for grade level or academic achievement. In some instances, upper-level students who were repeating the course as a result of poor prior performance during their freshman or sophomore years were matched with students in their second or third semesters who were on the Dean's List. It would be interesting to investigate whether deliberately assigning more successful students to work with less successful students (almost in a peer tutoring framework) would result in qualitatively different feedback being provided, or in quantitatively different exam results. However, faculty members wishing to pursue such a course should be cautioned that the process may not be equally valuable for both partners. The more successful student may be disadvantaged and, given the findings that Hacker et al (2000) outline with regard to more successful students being able to self-assess more accurately from the outset, the benefits reaped through the process may be distinctly one-sided.

A second modification that could be investigated in a future study involves the timing of the preliminary self-assessment component of the exam autopsy. The pilot project implemented a post-graded exam start time; however, it may be worth exploring whether asking students to conduct their preliminary self-assessment immediately following the exam, rather than waiting until they see their actual grade, would result in the same degree of statistical significance. In light of what the research literature presents about students' faulty predictions (and postdictions) of exam performance, it could be very telling to read what they believe about how well they prepared for the exam before they see their actual grade. Then, in their final reflective self-assessment, they could address an additional prompt that inquired about their actual performance versus their perceived performance.

Finally, research needs to be undertaken to examine whether the exam autopsy model as presented in the pilot project is equally effective for different types of tests (i.e., short answer or essay exams, where greater emphasis is placed on critical thinking and writing ability) and, indeed, for different types of assignments (i.e., lab reports, research papers, oral presentations, etc.). One possible modification for exams or assignments that are not objective in nature would be to afford students the opportunity to revise and resubmit their work following the autopsy process. This could shed light on whether or not students truly are receptive to feedback and successfully incorporate the suggestions that are presented to them.

Faculty members seeking to implement the exam autopsy model in their classes should be cautioned that the process is a time-consuming one. It takes time to introduce the idea of metacognition in class, and to engage students in the process so that they want to take it seriously and provide honest evaluations. It takes time to review each student's preliminary self-assessment and to offer concrete suggestions based upon each individual comment. It takes time to meet

with each individual student outside of class to elaborate verbally on those written comments. It takes time to present the do's and don'ts of effective peer feedback, and to allow for verbal interactions between students as well as opportunities for writing out thoughts. All of these activities certainly do dramatically reduce the in-class time available for covering content, and some faculty members may be resistant to the idea of forfeiting precious contact time for what they may view as a less important learning goal. Indeed, when the exam autopsy process was introduced in various faculty development forums as a follow-up to the current study, and informal interviews were held with colleagues in different departments and at other institutions to determine whether they would be willing to experiment with using this approach, the first and foremost concern that faculty members expressed was "losing time." Promoting student accountability and self-regulated learning was lauded as a priority, yet faculty members lamented that there was so much material they needed to cover in a given semester, they simply could not fathom how that would be accomplished if so much time were sacrificed for the sake of this process.

In response to that concern, faculty members may wish to move some of the assessment activities associated with the exam autopsy out of class, in an online form. That is certainly one option. Yet it should be reiterated that, in its current form, this appears to be a promising integrative assessment model for promoting metacognition and reflective practice in students, which could, in turn, result in greater transfer of learning and enhanced academic achievement. The triangulation of three distinct sources of feedback (from self, instructor, and peer) and the opportunity to reflect on how closely aligned (or how far apart) these may be means that student misperceptions of their own performance and ability can be corrected. Many students rely on particular study strategies (such as highlighting or skimming the text, or cramming the night before) because these are familiar, convenient, and comfortable. If their misperceptions of their own abilities may be attributable to ignorance, so too can their knowledge of how to study be rooted in same; perhaps they study poorly because that is the only way they know how to study. If and when they are exposed to faculty members and fellow students who introduce alternative approaches, and who frame these approaches in a non-threatening, supportive manner, then they may be more likely to seize upon the opportunity to

try something new. Surely exposing students to the principles and practices necessary for lifelong learning, and initiating a shift from a victim to creator mindset which will have ramifications for critical behavioral changes, is a priority.

## REFERENCES

- Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How Learning Works: Seven Research-Based Principles for Smart Teaching*. San Francisco, CA: Jossey-Bass.
- Bercher, D. A. (2012). Self-monitoring tools and student academic success: When perception matches reality. *Journal of College Science Teaching*, 41 (5), 26-32.
- Butler, D. (1997, March). *The roles of goal setting and self-monitoring in students' self-regulated engagement of tasks*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Delclos, V. R., & Harrington, C. (1991.) Effects of strategy monitoring and proactive instruction on children's problem-solving performance. *Journal of Educational Psychology* 83 (1), 35-42.
- Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation*, 14 (7). Retrieved from <http://pareonline.net/pdf/v14n7.pdf>
- Dunning, D. (2007). *Self-Insight: Roadblocks and Detours on the Path to Knowing Thyself*. New York, NY: Taylor & Francis.
- Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 12 (3), 83-87.
- Downing, S. (2014). *On Course: Strategies for Creating Success in College and in Life*. Boston, MA: Cengage.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new era of cognitive-developmental inquiry. *American Psychologist*, 34 (10), 906-911.
- Gielen, M., & De Wever, B. (2015). Structuring peer assessment: Comparing the impact of the degree of structure on peer feedback content. *Computers in Human Behavior*, 52, 315-325.
- Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, 92, 160-170.
- Iwamoto, D. H., Hargis, J., Bordner, R., & Chandler, P. (2017). Self-regulated learning as a critical attribute for successful teaching and learning. *International Journal for the Scholarship of Teaching and Learning*, 11 (2), 1-10.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77 (6), 1121-1134.
- National Research Council. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.
- Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research*. San Francisco, CA: Jossey-Bass.
- Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeider (Eds.), *Handbook of Self-Regulation* (pp. 451-502). San Diego, CA: San Diego Academic Press.
- Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research, & Evaluation*, 11 (10). Retrieved from <http://pareonline.net/pdf/v11n10.pdf>
- Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68 (3), 249-276.
- Van den Berg, I., Admiraal, W., & Pilot, A. (2006). Peer assessment in university teaching: Evaluating seven course designs. *Assessment & Evaluation in Higher Education*, 31 (1), 19-36.
- Warkentin, R. W., & Bol, L. (1997, April). *Assessing college students' self-directed studying using self-reports of test preparation*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman, & D. H. Schunk (Eds.), *Self-Regulated Learning and Academic Achievement*, 2<sup>nd</sup> Ed. (pp. 1-38). Hillsdale, NJ: Erlbaum.
- Zimmerman, B. J., & Schunk, D. H. (2011). *Handbook of Self-Regulation of Learning and Performance*. New York, NY: Routledge.