Comparing Training Approaches for Technological Skill Development in Introductory Statistics Courses

1. INTRODUCTION

Technology is an inseparable part of the modern statistics course (Gould, 2010). Technology use in statistics education has been on the rise for the last couple of decades coinciding with recommendations laid out by the statistics education reform movement (American Statistical Association, 2005; Cobb, 1992). Recent survey reports surfacing from the U.S. suggest that up to 76% of statistics courses regularly use technology (Hassad, 2012). This has risen from an estimate of 50% identified in a previous U.S. survey (Garfield et al., 2002). This upward trend was predicted by Garfield et al.'s 2002 report, which found that 82–90% of statistics instructors anticipated making further changes to their courses involving the use of technology. Statistics education reform has strongly advocated the use of technology in introductory statistics courses as a way of fostering students' conceptual understanding and shifting the focus of courses away from computation. The types of technology that are utilized in introductory courses vary but common examples include statistical packages, educational software, spreadsheets, applets, graphing calculators, multimedia material, and data repositories (Chance et al., 2007).

The overall attitude towards the use of technology in statistics education has been to focus on "the content, and not the tool" (Chance et al., 2007, p. 4), but recently, some instructors have challenged this view. These instructors cite that the changing nature of statistical practice and unprecedented access to data will have a profound impact on statistics education (Gould, 2010; Nolan and Temple Lang, 2010a; Nolan and Temple Lang, 2010b). As Gould (2010) explains, the ability to use statistical technology is now a fundamental component of statistical literacy, not a mere "hurdle" (p. 309) suggested by the prevailing attitude. The best example to illustrate this point is the ability to operate statistical software packages (e.g. *SPSS*, *Minitab*, *SAS*, *Stata*, and *R*). The ability to use a statistical package is a vital skill that students must develop if they are to become statistically literate. Without this technological skill, students cannot meaningfully and practically analyze complex real-world data. In many cases, implementing modern statistical methods is completely impractical without the aid of a statistical package (e.g. creating plots, running simulations, statistical modeling, and bootstrapping). The absence of a discourse on the development of these types of technological skills suggest that most instructors assume students will just "pick up" (Gould, 2010) these skills and carry them throughout their career. Sadly, the opposite is most likely true.

If technological skills, such as statistical package skills, are fundamental to modern notions of statistical literacy, these skills need to be fostered in introductory statistics courses. The statistics education literature has fallen behind on understanding how these skills can be developed. Many fundamental questions must be addressed. How do students learn to use technology? What are the barriers to developing technological skills? How can instructors better foster students' technological skills? Many of these questions have been addressed in the general software training literature (e.g. organizational training for word processors, email, internet use, spreadsheets, and presentation software). While statistics education can draw from this knowledge base, the unique context of the introductory statistics course is likely to present many challenges. For one, the ability to use statistics technology is likely to be highly dependent on statistical knowledge. Separating students' technological skills from their conceptual understanding of statistics presents a major challenge (Baglin et al., 2011; Baglin et al., 2012a; Baglin et al., 2012b), which makes understanding the development of these skills difficult. The use of theoretical model may help focus research efforts.

Baglin (2012) proposed Kanfer and Ackerman's (1989) integrative model of skill acquisition as a useful framework for understanding the development of technological skills in statistics education. Kanfer and Ackerman's model explains technological skill acquisition by integrating students' cognitive ability and motivation within an information processing framework. The model begins with the idea that all students bring with them a level of cognitive ability. The model predicts that the more cognitive ability a student has, the better their training performance and subsequent training transfer.

Training transfer is defined as the ability to transfer skills outside of the training environment (Hesketh, 1997). Kanfer and Ackerman's model also incorporates students' motivation. The more motivated students are to learn during training, the more likely they are to commit the required cognitive processes to a task. Motivation can compensate for or compound the effect of cognitive ability.

While there are many models of motivation, the literature in this area discusses the concept of *perceived performance utility* (Keith et al., 2010). High performance utility is evident in students who value developing statistical package skills. Because they value or see utility in training, these students are highly motivated to engage. Kanfer and Ackerman's model predicts that student motivation and cognitive ability interact to determine students' engagement in training, which impacts their subsequent development of technological skills (i.e. training transfer).

Previous studies suggest that training transfer can be improved by using different training approaches (Bell et al., 2008). A training approach is a theoretical framework that guides the design and delivery of technology training. This can be contrasted with training delivery methods (e.g. computer laboratory sessions, in-class demonstrations, and self-guided modules). A large body of research that has looked at general software training has found that *active-exploratory training* (A-ET) approaches appear to have improved training transfer outcomes when compared to traditional *guided training* (GT) approaches (Keith et al., 2008; Bell et al., 2008; Chillarege et al., 2003; Keith et al., 2010; Frese et al., 1991; Heimbeck et al., 2003; Nordstrom et al., 1998; Wood et al., 2000). GT is founded on the programmed learning method developed by the famous behaviorist B. F. Skinner (1968). GT views the student as a passive participant during training. The student is presented with step-by-step, comprehensive and explicit instructions that guide them through learning to operate a statistical package. The GT approach is error-avoidant, i.e. errors are viewed as a nonproductive waste of time. Students' skills are developed through repeated practice where operational errors are minimized. On the other-hand, in A-ET the student is presented with minimal instruction that engages them in actively-exploring the statistical package (Bell et al., 2008). As comprehensive instructions are avoided, the student becomes an active participant in the development of their skills.

The most successful type of A-ET, *error-management training* (EMT), goes one step further. EMT pays special attention to the function of errors made during training. As students actively-explore the statistical package with minimal instruction, they will invariably commit errors. According to EMT, errors are encouraged as they lead to a deeper understanding of a software system, the know-how to avoid errors, the ability to explore new features of the package and the ability to deal with errors when they occur (Frese et al., 1991). To help deal with the typical negative emotions experienced after making an error, EMT incorporates emotional control strategies. Heuristics are presented to students during training, such as "*Errors are a natural part of learning. They point out what you can still learn!*" (Dormann et al., 1994, p. 368). These heuristics are designed to help students view errors in a positive light. They are delivered to students in training content and recited by trainers (e.g. tutors) present during training sessions.

Research comparing A-ET approaches to GT approaches have differentiated between two major types of training transfer outcomes, adaptive transfer and analogical transfer. *Adaptive transfer* is demonstrated in a student's ability to adapt limited training skills in order to confront novel situations outside of training (Keith et al., 2010). For example, a student may have learned how to conduct a two-sample *t*-test in a statistical package. Suppose they learn about one-way ANOVA in another course and want to use the statistical package to run this procedure. Adaptive transfer would be evident if the student could adapt their skills of conducting two-sample *t*-tests to figure out how to operate the statistical package to perform the one-way ANOVA. Another example of adaptive transfer would be a student transferring their knowledge of one statistical package to learn a different statistical package. Adaptive transfer is the most desirable outcome of training as it promotes sustainable learning beyond the brief experience afforded by most training. Training should provide students with a foundation that they can continue to adapt and build upon outside of the training environment. The other type of training transfer, *analogical transfer*, is simply the ability to transfer the same skills covered in training. For example, if a student completed the topic of correlation in training, analogical transfer is evident if the student can perform correlation outside of training (Keith et al. 2010).

A meta-analysis that combined the results of 24 studies assessing the effect of EMT found an overall significant and positive effect over GT (Keith et al., 2008). Keith and Frese combined the results of experiments looking at general software training for simulation, word processing, databases, presentations, spreadsheets, e-mail, web browsers, and programming languages. The outcomes of this analysis found that EMT was significantly superior to GT for promoting adaptive transfer, and, to a lesser extent, analogical transfer. The study also found that the two core components of EMT, active-exploration and error-framing, contributed unique training effects suggesting that EMT is more effective than A-ET alone. Keith and Frese concluded that their results suggest that EMT is the preferred method of training when adaptive transfer is the goal.

The development of self-regulatory skills has been posited to explain the superiority of EMT. According to Keith et al. (2010), A-ET approaches, such as EMT, work by developing students' self-regulatory skills. Self-regulatory skills in a training context can be defined as a student's ability to guide their engagement in training activities by controlling cognition, mood, behavior and focus (Karoly, 1993, p. 25). This involves both metacognition and emotional control. Ford et al. (1998) define metacognition as a student's ability to exert "control over his or her cognitions" (p. 220) by planning, monitoring and evaluating task performance (Brown et al., 1983). Emotional control can be defined as "the use of self-regulatory processes to keep performance anxiety and other negative emotional reactions (e.g. worry) at bay during task engagement" (Kanfer et al., 1996, p. 186). As Keith et al. explain, minimal instruction promotes active-exploration, which requires students to practice metacognitive skills. Students must plan, monitor and evaluate how they are progressing through the training activities. GT, however, creates a passive training environment where students progress by following instructions. They do not engage at the same level of metacognitive activity required by EMT. Students in EMT are also required to develop emotional control strategies to deal with negative emotions created by errors. The EMT approach achieves this by creating an environment where students become habituated to making errors and by helping students realize their positive functions. Students in a GT approach avoid errors and become accustomed to the artificial use of guided instructions. They are not presented with the opportunity to develop emotional control strategies that are required when transferring skills in real-world situations outside of a "safe" error-free training environment.

Enhancing the promise of the EMT approach, Keith et al. (2010) found that A-ET curbed the effect of low motivation and low cognitive ability on adaptive training transfer. Kanfer and Ackerman's model suggests that the efficacy of training can be reduced for students who lack motivation to develop statistical package skills and students who may have lower cognitive ability. Keith et al. found that participants' trained using EMT for presentation and word processing software exhibited no relationship between adaptive transfer and participants' motivation or cognitive ability. On the other hand, participants' adaptive transfer for the GT condition was correlated with the participants' willingness to learn and general cognitive ability. Keith et al. explains that participants in EMT developed their self-regulatory skills in training more so than participants in the GT condition. When these participants were required to approach novel (adaptive) transfer tasks that were not addressed during training sessions, these participants were able to draw upon metacognitive skills and emotional control strategies developed during training. Participants in the GT condition did not develop these skills and when the time came to complete the adaptive transfer tasks, the participants' performance was predicted by their training motivation and general cognitive ability. Not only has EMT been found to be superior to GT for adaptive transfer, but Keith et al. (2010) suggest that EMT may be the preferred training approach for diverse student populations that exhibit differences in motivation and cognitive ability. However, the degree to which these findings extrapolate to the development of technology skills for statistical packages remains in question.

Few studies have looked at the effect of training approaches on the development of statistical package skills with the exception of Dormann and Frese (1994) and Baglin and Da Costa (2012a). Dormann and Frese randomly assigned 30 psychology students to be trained to use the statistical package *SPSS*. Participants completed a single training session that lasted two hours. In the following hour, training transfer was evaluated. The study did not specifically measure adaptive transfer, but instead, divided tasks between easy, moderate and difficult. The results indicated that participants in the EMT condition performed significantly better on measures of moderate and difficult training transfer tasks. Yet, this experiment had many limitations. The experiment was not based on an ecologically valid model of statistical package training. Two hours of training cannot be compared to a full semester program and immediate evaluation of training transfer fails to consider skill retention. The study also had a very small sample size and did not explicitly measure adaptive transfer. Dormann and Frese's findings required follow-up in a real introductory statistic course.

Baglin and Da Costa (2012a) compared EMT to GT in an introductory statistics course for psychology students. The study randomly allocated 100 students to six *SPSS* computer laboratory training sessions delivered using either a GT or EMT approach. Training sessions lasted one hour and were delivered fortnightly across the semester. Training transfer was measured using self-assessment exercises in the final weeks of the semester. In the second semester, 79 of the original participants completed the same follow-up self-assessment exercises from semester one. The results of the study found no statistically significant difference in measures of adaptive training transfer between the EMT and GT approaches. However, there were many limitations to this study that prevented clear conclusions. As Baglin and Da Costa report, embedding a randomized-experiment into a real statistics course and maintaining internal validity was challenging. Un-blinded participants, technical issues throughout the semester, limited computer laboratory availability, student non-compliance, and low-student engagement with self-assessment exercises were raised as limitations. Baglin and Da Costa also looked at manipulation checks and found that the EMT approach may have been invalidated by time constraints imposed on students. As training transfer was highly correlated with participant's end of semester exam scores, the authors also questioned the validity of their self-assessment transfer measures. Baglin and Da Costa suspected it may very well have been measuring knowledge of statistics instead of participants' statistical package skills.

Training transfer is not the only consideration that will impact the decision to use a particular training approach for the development of technological skills. Instructors must consider other factors such as training difficulty, anxiety and overall satisfaction. Training should also positively impact students' perceptions of self-efficacy. As Baglin and Da Costa (2012a) explain, instructors might be concerned that EMT may increase training difficulty leading to increased anxiety, lower perceptions of self-efficacy and lower overall training satisfaction. While Baglin and Da Costa failed to find evidence of such an effect, the same limitations outlined previously require reevaluation of this assertion.

The findings of Dormann and Frese (1994) and Baglin and Da Costa (2012) warrant further investigation into the effect of different training approaches for the development of technological skills in introductory statistics courses. The aim of this study was to build upon this previous research comparing GT and EMT approaches by addressing the key limitations outlined by Baglin and Da Costa (2012). Specifically, this study improved the validity of the implementation of the EMT approach, increased overall training time across the semester, blinded participants to the nature of study, and developed an improved measure of adaptive transfer. This study opted for a quasi-experimental design due to practical and ethical issues imposed by implementing randomized studies in educational settings. While randomized studies are considered the gold standard for evaluating educational interventions, research suggests that quasi-experimental designs can provide reliable estimates of causal effects provided adjustment for known covariates has taken place (Shadish et al., 2008). Important and known covariates were measured and controlled for to improve the comparisons between training approaches. This study chose to focus only on adaptive transfer outcomes as these were considered the most pertinent outcome of statistical package training. It was hypothesized that EMT would lead to significantly better statistical package adaptive transfer skills. To explore the possible implications of using EMT over GT, measures of student self-efficacy, training satisfaction, training anxiety, and training difficulty were also compared.

2. METHOD

Participants were recruited from a two-semester introductory statistics course for psychology students that ran concurrently across two campuses, A and B. The psychology programs at the two campuses were exactly the same. However, as the location of Campus B attracts a greater number of students the academic entrance requirements for enrollment in Campus B were higher than Campus A. The first semester course topics included exploratory data analysis, statistical inference for categorical variables and correlation. In the second semester, the course introduced inference of means and regression. Campus A was arbitrarily designated the EMT approach and Campus B the GT approach. The course was delivered across both campuses by the second author of this article. This author only presented the lecture content and had no interaction or involvement with students during training sessions. Only tutors, who were independent from the study, interacted with students during training.

Campus A had 41 students enrolled of which 35 (85%) consented to participate in the study. Campus B had 127 students enrolled of which 93 (73%) consented to participate. By the end of the study, 34 (97%) and 81 (87%) participants completed the requirements of the study from Campus A and B respectively (N = 115). Campus A had a mean age of 22.32 years (SD = 6.95) with 24 (74%) females. Campus B had a slightly lower mean age of 20.20 years (SD = 3.21) with 55 (68%) being female. Participants were asked at the beginning of the study if they had been previously trained to use the statistical package *SPSS*. There were two (6%) participants from Campus A and nine (11%) participants from Campus B who reported being previously trained.

2.1 Measures

Due to the quasi-experimental design of this study, it was important to control for pre-existing differences between the training approaches, which may explain variability in training transfer measures. Statistically controlling for these variables allowed for a better estimation of the association between training approaches and training transfer. Based on Kanfer and Ackerman's model, a student's cognitive ability will explain a large degree of the variability in training transfer outcome measures. Cognitive ability is a broad general construct that requires specialized testing (e.g., IQ testing) that was beyond the scope of this study. Therefore, a substitute variable for controlling for this effect was needed. A student's knowledge of statistics, as measured by average test and exam performance across the semester was chosen for this purpose. This was calculated by averaging the student's grade percentages across test 1, test 2 and the final exam. (Note: If a student missed any assessment, they received the average of the assessment they had completed.) While statistic exams scores have been found to be very weakly correlated with intelligence (e.g. Furnham & Chamorro-Premuzic, 2004), they do provide a more relevant way of controlling for the effect of student ability on training transfer. As previous research suggests, statistical knowledge is related to statistical package training transfer, suggesting that a student's knowledge of statistics will impact their development of statistical package skills (Baglin et al., 2012b; Baglin et al., 2012a). Therefore, to disentangle the effect of training approaches on adaptive training transfer, statistical knowledge was controlled for between training approaches.

Students' motivation to learn statistical packages was also taken into account. While there are many models of motivation that could be considered, this study took a direct approach similar to Keith et al. (2010). This involved measuring students' self-reported perceived performance utility. Statistical package performance utility was defined as the extent to which a student viewed *SPSS* as being useful technology for doing statistics. A scale to measure statistical package performance utility was created by adapting items from the *Questionnaire for the Content-Differentiated Assessment of Attitudes toward the Computer* by Richter, Naumann, and Groeben (2000). The subscale was originally designed to measure the extent to which a participant viewed a computer as being a valuable tool in everyday life. An example of an adapted item is "*SPSS will be a useful tool for doing my statistical analysis.*" The seven items that made up this scale were rated on a seven-point Likert-type scale ranging from strongly disagree (1) to strongly agree (7). Participants rated these items in a pre-training questionnaire given to students in the first lecture following an in-class demonstration of *SPSS*. Scores were averaged to get an overall performance utility score. The original items from Richter et al. had evidence of good psychometric properties. However, these metrics were reanalyzed following adaptation for the purpose of this study. Using a Principal Component Analysis (PCA) and selecting components with eigenvalues greater than one, a single component was extracted that explained 62% of the variation in responses. Internal consistency of the scale was found to be high (Cronbach's α = .88).

Students' progress through the training was recorded by counting the number of training sessions each student had completed up to one week prior to assessment of adaptive training transfer. As there were a total of ten training sessions, scores on this covariate could range from 0 to 10. The post-training questionnaire also asked participants to self-report the number of training sessions that they completed outside of their designated training session times. This variable was included to take into account for possible differences between the campuses that related to how the students completed the training. This was important to include as training was available online outside of scheduled training times. As this measure was self-reported on the post-training questionnaire, 32/93 (34.8%) participants in GT and 3/32 (9.4%) participants in the EMT approach were missing data. In the post-training questionnaire, participants were also asked if they had personal access to the statistical package. This was important to take into account as students with personal access may systematically differ from students who could only access the package on campus. Gender and age were also recorded.

2.2 Manipulation Checks

In line with previous studies, it was important to evaluate the validity of the imposed training approaches. Baglin and Da Costa (2012a) reported limitations with the manipulation of training approaches as a possible explanation for their null findings. Therefore, manipulation checks were included as a measure of internal validity. Self-reported measures of metacognitive activity, emotional control, exploratory behavior, the use of instructions and error orientation during training were included in the post-training questionnaire. All measures were rated on a seven-point Likert-type scale, ranging from (1) strongly disagree to (7) strongly agree. Scale scores were calculated by averaging participants' responses across items. Items that needed to be reverse-coded were reversed prior to averaging.

Metacognitive Activity. Metacognition was measured using 12 items adapted from Ford et al. (1998). The items required participants to self-report their level of metacognitive activity (i.e. monitoring, planning, and revising) exercised during training. An example of an item is "*I tried to monitor closely the statistical procedures in SPSS where I needed the most practice.*" Using a PCA and selecting components with an eigenvalue greater than one, a single component was extracted that explained 48% of the variability in responses. Internal consistency of the scale was found to be high (Cronbach's $\alpha = .89$).

Students' rated the degree to which they exercised emotional control during training using eight items originally adapted from Keith and Frese (2005) by Baglin and Da Costa (2012a). An example of an item is "*When difficulties arose during computer labs I was able to focus all my attention.*" A PCA of the eight items extracted a single component that explained 46% of the variability in responses. The internal consistency of responses to the scale was high (Cronbach's $\alpha = .82$).

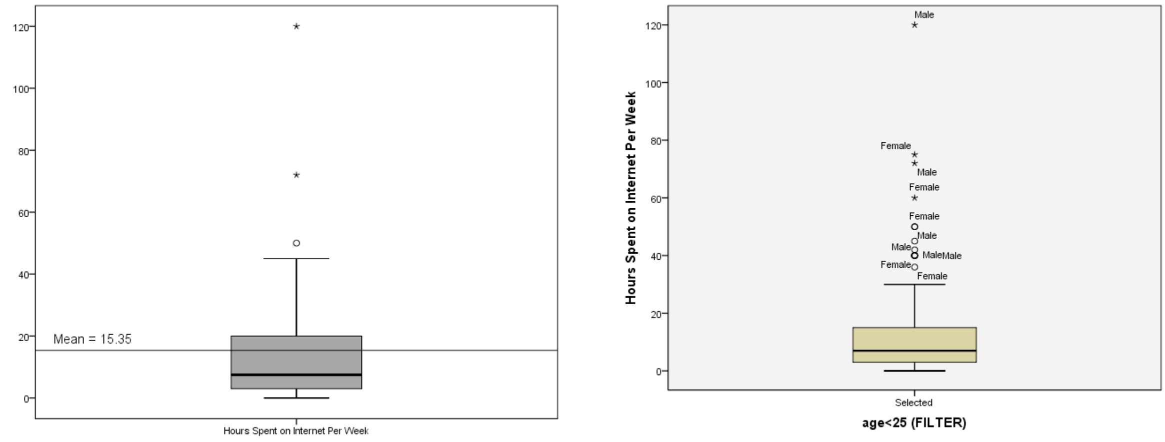
Participants' attitudes towards errors made during training were measured using the *Error Strain* and *Learning from Errors* subscales of the *Error Orientation Questionnaire* (EOQ, Rybowiak et al., 1999). Baglin et al. (2012a) adapted these items to refer to errors made during statistical package training. The *Error Strain* subscale measured the degree to which participants felt negative emotions when making errors (e.g. "*I was afraid of making errors when learning to use SPSS*") using five items and the *Learning from Errors* subscale measured the degree to which participants viewed errors as being a valuable learning experience (e.g. "*From my errors, I have learned a lot about how to work with SPSS*") using four items. A PCA, which forced the extraction of two components, confirmed the two-subscale structure of the EOQ. Both components explained a total of 62% of the variation in responses. Internal consistency of the subscales as measured by Cronbach's alpha were $\alpha = .79$ and $\alpha = .82$ for *Error Strain* and *Learning from Errors* respectively.

The degree to which students participated in exploratory or guided behavior during training was measured using six self-reported items borrowed and adapted from Bell and Kozlowski (2008). Three of these items related to exploratory behavior consistent with EMT, (e.g. "*I tried to discover how to operate SPSS without any instruction*"). The other three items measured students' behavior consistent with GT (e.g. using instructions, modeling others and seeking assistance from tutors). An example of an item is "*When I was unsure about how to complete a task in SPSS, I would immediately ask the tutor/or a friend for help.*" To aid the comparison with the Baglin and Da Costa (2012a) study, the mean rating of individual items were considered when checking the validity of the training approaches.

Adaptive Training Transfer. An *SPSS* certification task was used to measure adaptive transfer. (Note: Analogical transfer was not considered because adaptive transfer was the goal of training and the most important training outcome.) The certification task was scheduled for the final week of the semester and participation in the task contributed to 5% of a student's grade. The certification task was included to increase students' engagement in training during the semester. Baglin and Da Costa (2012) raised poor student engagement as an issue for measuring training transfer. The certification task was designed to increase student engagement by making students aware of the activity early in the semester, by making the task sound official, and attributing a higher grade to its completion than regular training. The task lasted one hour and was completed under exam-like conditions (e.g., no talking, no assistance). However, students were allowed to bring a copy of the course's *SPSS* quick reference guide (described in Section 2.3). The certification task presented students with six exercises. For each exercise, *SPSS* output was presented on a printed handout. Using a data file provided to them, the students had to replicate the output using *SPSS* for each exercise as closely as possible. The closer the student replicated the output, the higher their training transfer. The first two tasks were designed to be very simple and were not included in adaptive transfer scores. The remaining four tasks were designed to measure adaptive transfer and were scored out of 32. The exercises were adaptive because students had to replicate output that required them to adapt their training knowledge. This involved being able to link multiple procedures together that were treated separately during training (e.g. segregate a data file, filter out specific cases and create a plot) as well as manipulate and edit output (e.g. adding labels, reference lines and markers) in ways in which training did not cover.

Students were instructed to export their single closest replication of each exercise to a word processing document and upload it to an online submission site before leaving the certification session. There were three versions of the certification task worksheets (A, B, C). Appendix A contains a complete copy of version A. Each version was slightly different to prevent students' copying from their neighbors. A grading rubric was developed to identify key elements of each exercise that indicated the student had successfully adapted their skills (see Appendix B). These key elements were scored higher than other elements of the output that didn't require adaptation. An example of one of these exercises and the marking breakdown is shown in Figure 1. All student attempts were labeled using student numbers to prevent the lead researcher, who completed all grading, from associating student names, memorized from class lists throughout the semester, with the different approaches/campuses. Attempts from each training approach/campus were mixed and graded together. This was done to blind the lead researcher as to which training approach/campus each attempt belonged to. For student feedback purposes, participants were given a grade, 0, 1, 2 or 3, which reflected their performance on the task. Students who scored 0–1 were given the opportunity to complete further training between semesters to brush up on their skills as SPSS would be used throughout the second semester.

Other Training Outcomes. Besides training transfer, it was also important to consider other training outcomes that may have an impact on students and instructors. This study considered the association between training approaches and students' perceptions of statistical package self-efficacy, overall training difficulty, and training satisfaction. Students' perceptions of the difficulty, anxiety experienced, and level of training preparedness for the certification task were also evaluated. When giving their responses to the end of semester post-training questionnaire participants were asked to rate the overall difficulty and satisfaction of training on a seven-point Likert scale, ranging from (1) very easy/not at all satisfied to (7) very difficult/very satisfied respectively. On the same questionnaire, participants were also asked to rate their level of statistical package self-efficacy. Statistical package self-efficacy was defined as a participant's confidence in their ability to operate a statistical package after training. Three items from Finney and Schraw's (2003) *Current Statistics Self-Efficacy* (CSSE) scale were adapted for this purpose. Participants were required to rate their level of confidence in their current ability to use *SPSS* for generating descriptive statistics, graphical displays and statistical inference. An example of an item is "*To use the statistical package to conduct statistical inference (e.g. generate p-values).*" A similar seven-point Likert scale, ranging from (1) no confidence at all to (7) complete confidence, was used. Scores for the three items were averaged to form a single self-efficacy score (Cronbach's $\alpha = .78$).



a) Target output to be replicated
b) Example of student attempt

Figure 1. In this certification exercise the student was required to replicate a simple boxplot. However, to do this they were required to first select all males under the age of 25. This required the use of a filter with two conditions. Training only covered the use of simple filters and therefore, students had to adapt their knowledge to add another condition. This exercise also required the removal of outlier's labels, the insertion of a reference line at the mean, and a label for the value of the mean. Once again, these plotting options were not explicitly covered in training and required students to adapt their knowledge of basic plotting properties. In this attempt the student got 1 point for creating a boxplot, but failed to apply the correct filter (2 points), add a reference line and label for the mean (2 points) and remove the labels for the outliers (2 points). Thus, the adaptive components of each exercise were weighted higher.

Participants rated their level of anxiety that they experienced during training using four items originally adapted by Baglin and Da Costa (2012) from the *Tension-pressure* dimension scale of the *Intrinsic Motivation Inventory* created by Deci and Ryan and reported in McAuley, Duncan, and Tammen (1989). A sample item adapted by Baglin and Da Costa is "*I felt tense when training to use SPSS.*" Items were rated on a seven-point Likert-type scale, ranging from (1) strongly disagree to (7) strongly agree (7). Item ratings were averaged to obtain a scale score where higher scores equated to higher training anxiety. A PCA returned a single component that explained 55.34% of the variation in training anxiety scores. Internal consistency was calculated as Cronbach's $\alpha = .73$.

Before leaving the certification task session, participants were asked to rate the perceived difficulty of the certification task along with the level of anxiety they experienced and the degree to which they felt training had prepared them for the certification exercises. All questions were rated on a seven-point scale, similar to that used in the end of semester post-training questionnaire.

2.3 Training

Participants completed weekly one-hour statistical package training sessions in designated computer laboratories under the supervision of tutors. These sessions were designed to introduce students to the use of the statistical package *SPSS v. 20* as well as reinforce statistical concepts covered in lectures. The training was delivered using an online proprietary web-based assessment system called *WebLearn* (similar to *Blackboard*). Participants completed five training modules made up of a training and practice session (10 sessions in total). Training session introduced new *SPSS* procedures and practice sessions reinforced previous training material. Students completed the certification task in the final week of the semester. Completion of each laboratory session and the certification task contributed to a 20% (10 laboratory sessions = 15%, certification task = 5%) course participation grade. The module topics included the following: Introduction to *SPSS* (overview, entering data, editing variables, saving files, descriptive statistics, basic plots, editing plots, exporting output), The Basics of *SPSS* (revision from lab 1, boxplots, histograms, segregating and filtering data), Frequencies in *SPSS* (revision from lab 1 and 2, frequency tables, bar charts, recoding variables, and computing new variables), Crosstabs in *SPSS* (revision from lab 1, 2, and 3, cross-tabulations, Chi-square tests of association, clustered bar charts), and Correlation in *SPSS* (revision from lab 1, 2, 3 and 4, scatter plots, matrix scatter plots, and correlations). To help reinforce statistical concepts covered in the course, formative multiple-choice questions were embedded throughout laboratory sessions for both training approaches. These questions pre-empted statistical concepts to be covered in training to help facilitate the correct interpretation of *SPSS* output. For example, before students created cross-tabulations of two categorical variables, participants were presented with questions that required them to practice interpreting row and column percentages.

All training sessions were delivered online using *WebLearn*. The training sessions presented students with exercises that required them to learn to operate *SPSS*. Students either entered data or downloaded data files to use during the training and practice sessions. To confirm that the student had successfully operated the package, each exercise contained a question about the *SPSS* output generated. Students would enter their answer to receive immediate feedback on whether they had successfully completed the exercise. Each exercise could be attempted multiple times. To get their participation grades, students were required to attain 75% or above. Feedback for incorrect answers was provided in a form consistent with the training approach (described below). Both training approaches were provided with a copy of an *SPSS* quick reference guide. This guide listed and briefly described the features and procedures of *SPSS* that were covered throughout training. The guide was provided in response to previous course feedback. Electronic copies were linked to all training sessions.

EMT. Students in the EMT approach (Campus A) were presented with instructions at the beginning of training that established the conditions under which EMT operated. The instructions promoted active exploration and a positive attitude towards making errors. Students were told to expect to make errors and that these errors were a natural part of the learning process. Students were encouraged to try to correct any errors or solve any problems they had before seeking assistance from the tutors. At the beginning of each EMT session, students were provided with notes providing a minimal instructional overview of the features and procedures of *SPSS* that they would be training to use. These notes contained screenshots showing students how to access these procedures, but the screenshots were not linked with exercises, nor where there any step-by-step instructions provided. This met the criteria of minimal instruction as students needed to explore and adapt these features to complete their training exercises. Tutors were not permitted to guide students, but instead to encourage students to find solutions themselves. Throughout training, error-framing heuristics were presented to students above the exercises they were completing (e.g. "*Errors are a natural part of learning, they point out what you can still learn.*") These heuristics were provided to remind students of the positive function of errors. If a student got an exercise wrong, feedback was provided in the form of a positive error-framing heuristic as well as a hint designed to help them solve their error (e.g. "*Try playing around with the order of the variables entered into your plot*").

GT. Students in the GT approach (Campus B) were instructed to carefully follow the step-by-step instructions given to them and to avoid making errors where possible. If students made a mistake, they were told to read back through the instructions to identify their mistake. If they were uncertain, they could ask the tutor for guidance. In the GT approach, each exercise provided students with comprehensive step-by-step instructions and screenshots guiding the student through the entire exercise. Students were given automatic feedback from *WebLearn* telling them to re-try the steps when they made an error. Students would then be given another exercise to practice the procedure covered by the step-by-step instructions. The goal of the GT was to have students practicing the statistical package in an error-avoidant environment.

3. RESULTS

Data analysis comprised of the following three phases: validating training approaches, modeling adaptive transfer scores, and comparing training approaches on other outcomes. In order to assess training validity, mean ratings on manipulation check items were compared between training approaches using a series of independent sample *t*-tests. This was important as the correct manipulation of training approaches related directly to the internal validity of the study. Adaptive transfer scores were modeled using one-way analysis of covariance (ANCOVA). ANCOVA allowed the mean adaptive transfer scores to be compared between training approaches after controlling for the effect of training covariates. It was important to control for covariates in these models due to non-random allocation of participants. Due to some covariates containing a high proportion of missing values, multiple imputation techniques were used to estimate missing values. This aimed to reduce possible bias introduced by standard listwise deletion and improve the statistical power of the models. Finally, a series of independent sample *t*-tests were used to compare mean self-reported ratings on other training outcomes in order to explore the possible implications of implementing either of the training approaches.

3.1 Validating Training Approaches

In order to evaluate whether the training approaches had been conducted appropriately, mean student self-report ratings on metacognition, emotional control, learning from errors, error strain, guided training behavior and exploratory training behavior were compared using a series of independent sample *t*-tests (**Table 1**). As these manipulation tests were used to explore the validity of the training contexts as opposed to formal evaluation of the primary outcome of this study, no corrections for inflated type I errors were included. All tests were two-tailed and compared to an unadjusted significance level of 0.05. The results of these tests revealed that participants' mean ratings of the EMT approach were significantly different to the mean ratings of participants in the GT approach on items of active exploration, exploration without instructions, metacognition, operation without instruction, seeking assistance, and the use of step-by-step instructions. Participants in the EMT approach reported significantly higher mean self-reported ratings of exploratory behavior, metacognition, and operation without instructions. However, there were no significant differences on ratings of error strain, copying from other students, emotional control or learning from errors (**Table 1**).

Table 1
Descriptive Statistics and Independent Sample t-tests Comparing Mean Ratings of Manipulation Checks between Training Approaches

								95% *CI* of Difference	
Manipulation Variable		*M*	*SD*	*N*	*SEM*	*t*	*p*	Lower	Upper
Metacognition	GT	4.46	1.06	57	0.14	-2.14	.035	-0.87	-0.03
	EMT	4.91	0.73	32	0.13			-0.83	-0.07
Emotional Control	GT	5.44	0.97	57	0.13	2.08	.041	0.02	0.84
	EMT	5.01	0.86	32	0.15			0.03	0.82
Learning from Errors	GT	4.83	1.05	57	0.14	-1.47	.146	-0.83	0.12
	EMT	5.18	1.14	32	0.20			-0.84	0.14
Error Strain	GT	2.91	1.46	57	0.19	-0.97	.333	-0.89	0.31
	EMT	3.21	1.17	32	0.21			-0.86	0.27
Used step-by-step instructions	GT	6.21	1.45	57	0.19	4.34	<.001	0.72	1.95
	EMT	4.87	1.29	32	0.23			0.74	1.93
Copied other students	GT	2.19	1.55	57	0.21	-1.18	.240	-1.07	0.27
	EMT	2.59	1.50	32	0.27			-1.07	0.27
Immediately sought assistance	GT	4.63	2.02	57	0.27	2.02	.046	0.02	1.81
	EMT	3.72	2.08	32	0.37			0.00	1.82
Actively explored *SPSS*	GT	3.70	1.79	56	0.24	-2.67	.009	-1.73	-0.25
	EMT	4.69	1.45	32	0.26			-1.69	-0.29
Operate without instruction	GT	3.61	1.87	57	0.25	-4.10	<.001	-2.34	-0.81
	EMT	5.19	1.47	32	0.26			-2.29	-0.86
Explored without instruction	GT	2.98	1.70	57	0.22	-5.39	<.001	-2.51	-1.15
	EMT	4.81	1.20	32	0.21			-2.45	-1.21

3.2 Modeling Adaptive Transfer Scores

Before modeling adaptive transfer scores, the first step was to identify important covariates. Descriptive statistics and intercorrelations for covariates and adaptive transfer scores between training approaches are shown in Table 2. Covariates that were statistically significantly correlated with adaptive transfer scores were selected as covariates. Gender, personal access, training progress, and statistical knowledge were all significantly and positively correlated with adaptive transfer scores. The personal access variable contained a high degree of missing values, 32/93 (34.8%) for GT and 3/32 (9.4%) for EMT.

Adaptive transfer scores were modeled using one-way analysis of covariance (ANCOVA). ANCOVA allowed for the mean adaptive transfer scores to be compared between training approaches after controlling for the effects of gender, personal access, training progress and statistical knowledge. The first model employed traditional listwise deletion of cases with missing values present in the personal access covariate. While the overall model was statistically significant, $F(5, 83) = 8.93$, $p < .001$, $\eta^2 = .35$, $N_{GT} = 57$, $N_{EMT} = 32$, training approach was not a statistically significant predictor of adaptive training transfer scores, $F(1, 83) = 0.22$, $p = .64$, $\eta^2 = .003$. Personal access, $F(1, 83) = 9.34$, $p = .003$, $\eta^2 = .10$, and statistical knowledge, $F(1, 83) = 15.86$, $p < .001$, $\eta^2 = .16$ were both statistically significant covariates (see Table 3). Gender, $F(1, 83) = 3.80$, $p = .06$, $\eta^2 = .04$, and training progress, $F(1, 83) = 0.80$, $p = .37$, $\eta^2 = .01$, failed to reach statistical significance in the model suggesting that personal access and statistical knowledge better accounted for adaptive transfer scores.

In order to study the influence of the personal access covariate, a second model, which employed a more powerful set of covariates, was fitted. The second model was also statistically significant, $F(4, 110) = 10.4$, $p < .001$, $\eta^2 = .27$, $N_{GT} = 81$, $N_{EMT} = 34$, but did exhibit a lower partial η^2 indicating a higher degree of unexplained variance (Table 3). Once again, training approach was not statistically significant, $F(1, 110) = 0.91$ $p = .343$, $\eta^2 = .01$, but it did enter the model showing a slightly larger effect. With the removal of personal access, gender became statistically significant, $F(1, 110) = 5.02$, $p = .03$, $\eta^2 = .04$, and statistical knowledge remained in place as the strongest predictor, $F(1, 110) = 26.37$, $p < .001$, $\eta^2 = .19$. As per the initial model, training progress was not statistically significant, $F(1, 110) = 0.34$, $p = .56$, $\eta^2 = .00$.

A comparison of the two previous models suggested some important co-variation between adaptive transfer scores, personal access and gender. Given that personal access was highly correlated with adaptive transfer scores (see **Table 2**) and there was a large difference in the proportion of students with personal access between training approaches (40.6% EMT vs. 14% GT), both of the previous models suffered serious limitations. Model 1 was underpowered and possibly biased by the listwise removal of missing cases and Model 2 completely ignored the personal access covariate. Thus, a third model was fitted to overcome these limitations.

The third model used a multiple imputation method to estimate missing values for the personal access covariate. While the assumption behind this procedure states that missing values are required to be missing at random (MAR) or missing completely at random (MCAR), studies suggest multiple imputation performs quite favorably in situations where data are not missing at random (non-MAR, Shrive et al., 2006; Greenland et al., 1995). As Schafer (1997) explains, multivariate data sets that exhibit robust associations between variables provide a useful basis for imputing missing values, which aids in minimizing possible bias introduced by imputation of non-MAR values.

Multiple imputation was performed using the IBM *SPSS* Missing Values 19 package. All covariates and outcome variables were specified in the model and ten imputations were obtained. Parameters estimates for the ten imputations were pooled together and used to construct the third ANCOVA model (Table 3). The results of the ANCOVA using pooled parameter estimates from multiple imputations of missing values validated the results of model 1. Personal access, $p < .001$, and statistical knowledge, $p < .001$, were the only statistically significant predictors of adaptive training transfer. There was no evidence of a statistically significant effect for training approach, $p = .98$.

Table 2

Descriptive Statistics and Intercorrelations of Covariates and Adaptive Transfer

Variable		1.	2.	3.		4.	5.	6.	7.	8.	9.
1. Gender[1]		-	.109	.170		.093	.240*	.067	-.116	.085	.221*
2. Personal Access			-	.010		.355**	.215*	-.199	.241*	-.024	.296**
3. Previous Training Experience				-		.138	.128	-.160	-.003	.091	.068
4. Age						-	.165	-.013	.286**	.143	.055
5. Labs Completed Outside Training							-	-.322**	-.036	-.061	-.013
6. Training Progress								-	.059	.421**	.238*
7. Performance Utility									-	.079	.052
8. Statistical Knowledge										-	.485**
9. Adaptive Transfer											-
GT	*N*	26/81 Male	8/57 Yes	9/81 Yes	*M*	20.2	3.09	9.16	5.61	72.38	13.79
	%	32.1%	14.0%	11.1%	*SD*	3.21	3.4	1.84	0.97	13.65	6.97
					N	81	57	81	79	81	81
EMT	*N*	9/34 Male	13/32 Yes	2/34 Yes	*M*	22.32	4.41	7.88	5.94	66.25	13.24
	%	26.5%	40.6%	5.9%	*SD*	6.95	2.80	2.58	0.75	14.42	7.32
					N	34	32	34	34	34	34

[1]Gender: Females = 1, Males = 2

* $p < .05$, ** $p < .01$

Table 3: ANCOVA Model Parameters Predicting Adaptive Transfer

		1. Listwise deletion				
Parameters	*B*	95% *CI*	*SE*	*t*	*p*	η^2
Gender[1]	2.74	(-0.06, 5.54)	1.41	1.95	0.055	0.04
Personal Access	4.82	(1.69, 7.96)	1.58	3.06	0.003	0.10
Training Progress	0.30	(-0.37, 0.96)	0.33	0.90	0.373	0.01
Statistical Knowledge	0.21	(0.10, 0.31)	0.05	3.98	< .001	0.16
Training approach[2]	-0.67	(-3.52, 2.17)	1.43	-0.47	0.640	0.00
GT Adjusted Mean	13.49	(11.88, 15.10)	*N* = 57			
EMT Adjusted Mean	14.16	(11.96, 16.37)	*N* = 32			
		2. Personal access removed				
Gender[1]	2.79	(0.32, 5.26)	0.03	2.24	0.027	0.04
Training Progress	0.17	(-0.42, 0.77)	0.56	0.58	0.564	0.00
Statistical Knowledge	0.23	(0.14, 0.32)	0.00	5.14	< .001	0.19
Training approach[2]	-1.24	(-3.83, 1.34)	0.34	-0.95	0.343	0.01
GT Adjusted Mean	13.26	(11.90, 14.62)	*N* = 81			
EMT Adjusted Mean	14.50	(12.36, 16.64)	*N* = 34			
		3. Multiple imputation of missing values				
Gender[1]	2.20	(-0.16, 4.56)	1.20	1.83	0.067	
Personal Access	5.32	(2.17, 8.47)	1.60	3.33	0.001	
Training Progress	0.36	(-0.23, 0.94)	0.30	1.20	0.232	
Statistical Knowledge	0.21	(0.13, 0.30)	0.04	4.85	< .001	
Training approach[2]	0.03	(-2.51, 2.57)	1.30	0.03	0.980	
GT Adjusted Mean	13.64	(12.34, 14.93)	*N* = 81			
EMT Adjusted Mean	13.60	(11.53, 15.68)	*N* = 34			

[1] Females = 1, Males = 2,
[2] GT = 1, EMT = 0

3.3 Other Training Outcomes

Independent sample *t*-tests were used to compare mean self-reported ratings between training approaches on training difficulty, training satisfaction, training anxiety, and post-training self-efficacy (Table 4). All tests were two-tailed and were compared to an unadjusted significance level of 0.05. Correction of inflated type I error was not made since these tests were exploratory in nature. Mean self-reported ratings of participants' perceptions of the certification task's difficulty, anxiety and degree of preparedness were also analyzed (Table 4). Evidence of a statistically significant difference in mean ratings was found for training difficulty ($p < .001$) and satisfaction ($p = .016$). There was no evidence of statistically significant differences in participants' ratings of training anxiety ($p = .79$) and statistical package self-efficacy ($p = .67$). In terms of participants' perceptions of the certification task, there was no statistically significant evidence of any differences existing between participants' mean ratings of difficulty ($p = .492$), anxiety ($p = .525$), and preparedness ($p = .655$).

Table 4
Descriptive Statistics and Independent-sample t-tests Comparing Training approaches on Other Training Outcomes

Outcome		*M*	*SD*	*N*	*SE*	*t*	*p*	95% *CI* of Difference Lower	Upper
Training Difficulty	GT	3.30	1.21	57	0.16	0.001	0.001	-1.46	-0.40
	EMT	4.23	1.18	31	0.21				
Training Satisfaction	GT	5.19	1.30	57	0.17	0.016	0.016	0.14	1.31
	EMT	4.47	1.39	32	0.25				
Training Anxiety	GT	3.16	1.20	57	0.16	0.788	0.788	-0.60	0.46
	EMT	3.23	1.22	32	0.22				
Self-efficacy	GT	4.98	1.13	57	0.15	0.671	0.671	-0.55	0.35
	EMT	5.07	0.80	32	0.14				
CT Difficulty	GT	4.87	1.15	77	0.13	0.492	0.492	-0.30	0.62
	EMT	4.71	0.94	31	0.17				
CT Anxiety	GT	4.24	1.59	78	0.18	0.525	0.525	-0.85	0.44
	EMT	4.45	1.39	31	0.25				
CT Preparedness	GT	4.48	1.37	77	0.16	0.655	0.655	-0.43	0.68
	EMT	4.35	1.17	31	0.21				

CT = Certification Task

4. DISCUSSION

The aim of this study was to evaluate the effect of two training approaches for the development of technological skills in statistics education. This study specifically examined statistical package skills and how different training approaches might promote the development of sustainable outcomes, i.e. adaptive transfer. The EMT approach, a sub-type of active-exploratory training, was hypothesized to promote adaptive transfer above and beyond a conventional GT approach. The hypothesis of this study was based on the positive outcomes of previous research that has assessed adaptive transfer for general software skills (e.g. computer simulations, word processors, database searches, and spreadsheets, Keith et al., 2008; Bell et al., 2008; Chillarege et al., 2003; Frese et al., 1991; Heimbeck et al., 2003; Keith et al., 2010; Keith et al., 2005). However, after controlling for covariates, the results of this study found no statistical evidence of an association between the EMT approach and students' level of adaptive transfer. These results contradict an early experiment evaluating statistical package skills by Dormann and Frese (1994), but confirm the results of a recent experiment by Baglin and Da Costa (2012a).

The findings of the Dormann and Frese (1994) experiment suggested initial promise for EMT for statistical package skills. However, this experiment had many limitations that required further research. Their conclusions were limited by short-term follow-up, a small sample, one-off training sessions, and no deliberate attempt to measure adaptive transfer. The Baglin and Da Costa (2012a) experiment also had limitations. Due to significant constraints imposed on educational research, the Baglin and Da Costa experiment confronted issues with a short duration of training, un-blinded participants, questionable validity of training transfer measures, questionable student engagement during the evaluation of training transfer, and questionable validity of the imposed EMT approach. Hence, the aim of this study was to address these limitations.

The strengths of this study lie in its ecological validity (positioned within a real introductory statistics course), careful manipulation of training approaches, and improved validity of the evaluation of adaptive transfer for statistical package skills. Regardless, this study still had limitations. While randomized experiments are highly regarded for this type of evaluation, randomized protocols are notoriously challenging to implement effectively in an educational setting. Quasi-experimental designs provide a feasible compromise. However, due to non-randomization, the potential for systematic bias between training approaches is high. Fortunately, research suggests that quasi-experimental designs can provide reliable approximations to randomized experiments providing proper adjustment to known covariates has taken place (Shadish et al., 2008). This study was designed prospectively to control for known covariates in the statistical analysis. However, the degree to which this study has approximated a randomized study cannot be known.

There were a number of differences between the training approaches, or campuses, that were likely to impact on the development of adaptive transfer. As Kanfer and Ackerman's model suggests, the cognitive ability of trainees will have an effect on training performance and subsequent training transfer outcomes. While statistical knowledge is no substitute for a measure of general cognitive ability, it does provide insight into the academic and statistical ability of participants. The descriptive statistics show a difference of six percent on average statistical knowledge scores between training approaches/campuses. This highlights a key difference in the students' academic abilities between the two training approaches. This difference is further supported by national tertiary entrance requirements for undergraduate university programs. Campus A (EMT) and Campus B (GT) entrance scores where respectively 68 and 77 out of a theoretical 100. This suggests that students who performed better in their final years of secondary school were more attracted to Campus B even though they in are the same psychology programs run across different campuses. Fortunately, the adjustment for statistical knowledge does reduce the possibility of bias attributed to differences in students' academic ability.

Differences between the campuses that could not be controlled for were the class and laboratory session times. Campus A lectures and computer laboratory sessions were scheduled from midday to mid-afternoon, and Campus B were scheduled during the mornings. Anecdotally, previous students from Campus A have raised concerns about the scheduling of the statistics course in the late afternoon stating that they felt tired by the time they got into the computer laboratory sessions. Students had a clear preference for morning sessions. However, due to institutional constraints, the computer laboratory sessions could only be scheduled during the afternoon. This difference between the campuses could explain a number of the study's observations. It may explain why the overall perceived difficulty and satisfaction of training was lower for EMT/Campus A. There is no doubt that being tired would lower overall satisfaction and increase perceived difficulty. This may also explain why many of the Campus A participants reported completing training sessions outside of the scheduled times more frequently. Completing more training sessions outside of class would also explain why their average level of training progress was lower prior to the certification task. The structure and weekly progression of the scheduled laboratory sessions would be more likely to keep students up-to-date. Forcing students to attend the scheduled laboratory computer sessions would have been possible, but doing so would have violated the ecological nature of this study. It was important for these courses to allow students to access training sessions in their own time.

Comparison of mean ratings on the manipulation check items showed that participants in the EMT approach engaged in less guided instruction and more exploratory behavior. This was a vast improvement on the manipulation checks reported by Baglin and Da Costa (2012a). Surprisingly, however, the error-framing aspect of EMT was not validated. The addition of an error-framing element to active-exploratory training has been found to provide a unique effect above and beyond active-exploratory training alone (Keith et al., 2008). The absence of an error-framing effect may have reduced the overall effectiveness of EMT. This study suggests that encouraging and promoting errors as a beneficial aspect of training for statistical package skills might present a unique challenge. Given that most students come from educational settings where errors are viewed as failure and something to be avoided, one semester of training may not have been enough to change students' perceptions and attitudes towards making errors.

The certification task, which aimed to measure statistical package adaptive transfer skills, was an improvement on Baglin and Da Costa's (2012) self-assessment exercises. The certification tasks were designed to minimize the effect of statistical knowledge on operating the statistical package. While students still required a basic level of statistical knowledge to understand the output that was given, this dependency was reduced since they didn't have to make decisions about what statistical methods to use. The students could concentrate on demonstrating their ability to operate the statistical package. Anecdotally, student engagement during the certification task was reported to be high. The certification task was the only training session that was compulsory to attend in person. Tutors were present during these sessions to ensure exam conditions were imposed. Making the certification task worth 5/20 (25%) for the computer laboratory course participation grade ensured that students took the task seriously.

The ecological validity of the certification task measuring adaptive transfer must also be discussed. The internal validity of the tasks, at least on face value, would appear to be high as all students attempted the tasks under similar conditions and the tasks required students to demonstrate objective evidence of their ability to adapt their knowledge. However, the ecological, or real world, validity of these tasks requires further research and investigation. It's difficult to judge to what extent this ability can be captured in a formal assessment environment. Outside formal assessment settings, students may exhibit evidence of adaptive transfer in novel and unanticipated ways that were stifled by the controlled environment put in place during the certification task. This tension between internal and ecological validity when evaluating adaptive transfer is a limitation that will prove difficult to overcome.

Overall, this study failed to support the efficacy of EMT. Therefore, it's important to consider possible reasons that may explain the null finding, which contradict the consensus for general software training (Keith et al., 2008). One possible explanation that requires further investigation is the potential mediating effect of prior knowledge on EMT. The effectiveness of EMT is based on studies using technological skills that do not require specialized prior knowledge (e.g. word processors, presentation software, and spreadsheets, Keith et al., 2008). Technological skills for statistics may present a special case, as these skills are likely to be highly dependent on a trainee's knowledge of statistics. This would explain the difference between this study and the findings of Dormann and Frese (1994). Dormann and Frese used participants who had already completed introductory statistics courses and may have already developed the necessary knowledge to benefit from EMT. This study, however, trained students during the development of the required prior knowledge. These students may have missed out on the benefits of EMT as they were still coming to terms with understanding statistical concepts. Therefore, low prior statistical knowledge may mediate the effect of EMT. Future research should test this hypothesis by evaluating EMT on students possessing prior statistical knowledge.

This study confirmed a moderate relationship between training transfer and statistical knowledge identified in previous work (Baglin et al., 2012a; Baglin et al., 2012b). This relationship suggests that students who have a better understanding of statistical concepts tended to develop statistical package skills better than students with lower statistical knowledge. However, there is still a large degree of unexplained variance suggesting that many other factors may come into play. This study asked participants if they had personal access to the statistical package. Students with personal access tended to perform better on measures of adaptive training transfer even after controlling for participants' statistical knowledge, gender, and training progress. Personal access may have provided students with greater opportunity to practice and the ability to better integrate the statistical package into their regular repertoire of software. The large degree of missing values present for this variable required missing value imputation and therefore the estimated effects of this study are limited. Regardless, this finding suggests an interesting avenue for future research to evaluate the importance of personal access to technology on the development of technological skills. The results of this study suggest that access will likely play a greater role than the use of different training approaches.

Technological skills, such as the ability to operate statistical packages, are an important part of modern notions of statistical literacy. While the focus of statistics education is to teach the concepts, instructors can no longer ignore the importance of technological skills, especially, as students become increasingly more reliant on the technology. Statistics education research needs to play a key role in understanding how these types of skills interact in statistics courses and how these skills are best developed. This study found no association between the development of statistical package skills and two different types of training approaches, active-exploratory training and guided training. However, this study identified important areas for future research. The potential mediating effect of prior statistical knowledge on technological skill development requires further investigation. Statistical knowledge was indeed the most important predictor of adaptive transfer. The importance of personal access to technology may also prove to be an important determinant. Further research is needed to understand how these factors and many other undiscovered factors can be manipulated to foster students' development of technological skills in statistics education.

5. REFERENCES

American Statistical Association. (2005). *Guidelines for the assessment and instruction of statistics education*. Washington, DC: Author.

Baglin, J., and Da Costa, C. (2012a), "An Experimental Study Evaluating Error Management Training for Learning to Operate a Statistical Package in an Introductory Statistics Course: Is Less Guidance More?," *International Journal of Innovation in Science and Mathematics Education*, 20, 48–67. Available at http://ojs-prod.library.usyd.edu.au/index.php/CAL/article/view/5809/6529.

Baglin, J., and Da Costa, C. (2012b), "Students' Thoughts and Perceptions of Training to Use Statistical Packages in Introductory Statistics Courses: A Qualitative Study," in *Proceedings of the Eighth Australian Conference on Teaching Statistics, July 2012*, ed. H. MacGillivray and B. Phillips. Adelaide, South Australia.

Baglin, J., Da Costa, C., Ovens, M., and Bablas, V. (2011), "An Experimental Study Comparing Two Different Training Strategies on How to Use Statistical Software Packages in an Introductory Statistics Course," in *Proceedings of the Australian Conference on Science & Mathematics Education (17th Annual UniServe Science Conference), September 2011*, ed. M. Sharma, A. Yeung, T. Jenkins, E. Johnson, G. Rayner, and J. West, pp. 162 – 168. Melbourne, Australia.

Bell, B. S., and Kozlowski, S. W. J. (2008), "Active Learning: Effects of Core Training Design Elements on Self-regulatory Processes, Learning, and Adaptability," *Journal of Applied Psychology*, 93, 296–316.

Brown, A. L., Bransford, J. D., Ferrara, R. A., and Campione, J. C. (1983), "Learning, Remembering, and Understanding," in *Handbook of Child Psychology: Cognitive Development*, ed. P. Mussen, J. H. Flavell, and E. M. Markman, pp. 77–166. 4th ed. New York: Harvard University Press.

Chance, B. L., Ben-Zvi, D., Garfield, J. B., and Medina, E. (2007), "The Role of Technology in Improving Student Learning of Statistics," *Technology Innovations in Statistics Education*, 1. Available at http://escholarship.org/uc/item/8sd2t4rr.

Chillarege, K. A., Nordstrom, C. R., and Williams, K. B. (2003), "Learning from Our Mistakes: Error Management Training for Mature Learners," *Journal of Business and Psychology*, 17, 369–385.

Cobb, G. W. (1992), "Teaching statistics," in *Heeding the Call for Change: Suggestions for Curricular Action (MAA Notes No. 22)*, ed. L. A. Steen, pp. 3–43. Washington, DC: Mathematical Association of America.

Dormann, T., and Frese, M. (1994), "Error Training: Replication and the Function of Exploratory Behavior," *International Journal of Human-Computer Interaction*, 6, 365–372.

Finney, S. J., and Schraw, G. (2003), "Self-efficacy Beliefs in College Statistics Courses," *Contemporary Educational Psychology*, 28, 161–186.

Ford, J. K., Smith, E. M., Weissbein, D. A., Gully, S. M., and Salas, E. (1998), "Relationships of Goal Orientation, Metacognitive Activity, and Practice Strategies with Learning Outcomes and Transfer," *Journal of Applied Psychology*, 83, 218–233.

Frese, M., Brodbeck, F., Heinbokel, T., Mooser, C., Schleiffenbaum, E., and Thiemann, P. (1991), "Errors in Training Computer Skills: On the Positive Function of Errors," *Human-Computer Interaction*, 6, 77–93.

Furnham, A., and Chamorro-Premuzic, T. (2004), “Personality and Intelligence as Predictors of Statistics Examination Grades,” *Personality and Individual Differences*, 37, 943–955.

Garfield, J. B., Hogg, B., Schau, C., and Whittinghill, D. (2002), “First Courses in Statistical Science: The Status of Educational Reform Efforts,” *Journal of Statistics Education*, 10. Available at http://www.amstat.org/publications/jse/v10n2/garfield.html.

Gould, R. (2010), “Statistics and the Modern Student,” *International Statistical Review*, 78, 297–315.

Greenland, S., and Finkle, W. D. (1995), “A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses,” *American Journal of Epidemiology*, 142, 1255–1264.

Heimbeck, D., Frese, M., Sonnentag, S., and Keith, N. (2003), “Integrating Errors into the Training Process: The Function of Error Management Instructions and the Role of Goal Orientation,” *Personnel Psychology*, 56, 333–361.

Hesketh, B. (1997), “Dilemmas in Training for Transfer and Retention,” *Applied Psychology*, 46, 317–339.

Kanfer, R., and Ackerman, P. L. (1989), “Motivation and Cognitive Abilities: An Integrative/aptitude-treatment Interaction Approach to Skill Acquisition,” *Journal of Applied Psychology*, 74, 657–690.

Kanfer, R., Ackerman, P. L., and Heggestad, E. D. (1996), “Motivational Skills & Self-regulation for Learning: A Trait Perspective,” *Learning and Individual Differences*, 8, 185–209.

Karoly, P. (1993), “Mechanisms of Self-regulation: A Systems View,” *Annual Review of Psychology*, 44, 23–52.

Keith, N., and Frese, M. (2005), “Self-regulation in Error Management Training: Emotion Control and Metacognition as Mediators of Performance Effects.,” *Journal of Applied Psychology*, 90, 677–691.

Keith, N., and Frese, M. (2008), “Effectiveness of Error Management Training: A Meta-analysis.,” *Journal of Applied Psychology*, 93, 59–69.

Keith, N., Richter, T., and Naumann, J. (2010), “Active/exploratory Training Promotes Transfer Even in Learners with Low Motivation and Cognitive Ability,” *Applied Psychology: An International Review*, 59, 97–123.

McAuley, E., Duncan, T., and Tammen, V. V. (1989), “Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis,” *Research Quarterly for Exercise and Sport*, 60, 48–58.

Nolan, D., and Temple Lang, D. (2010a), “Integrating Computing and Data Technologies into the Statistics Curricula,” in *Proceedings of the Eighth International Conference on Teaching Statistics, July 2010*, ed. C. Reading. Ljubljana, Slovenia. Available at http://icots.net/8/cd/pdfs/invited/ICOTS8_9B2_NOLAN.pdf.

Nolan, D., and Temple Lang, D. (2010b), “Computing in the Statistics Curricula,” *The American Statistician*, 64, 97–107. Available at http://www.tandfonline.com/doi/abs/10.1198/tast.2010.09132.

Nordstrom, C. R., Wendland, D., and Williams, K. B. (1998), “To Err Is Human: An Examination of the Effectiveness of Error Management Training,” *Journal of Business and Psychology*, 12, 269–282.

Richter, T., Naumann, J., and Groeben, N. (2000), "Attitudes Toward the Computer: Construct Validation of an Instrument with Scales Differentiated by Content," *Computers in Human Behavior*, 16, 473–491.

Rybowiak, V., Garst, H., Frese, M., and Batinic, B. (1999), "Error Orientation Questionnaire (EOQ): Reliability, Validity, and Different Language Equivalence," *Jounrnal of Organizational Behavior*, 20, 527–547.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Shadish, W. R., Clark, M. H., and Steiner, P. M. (2008), "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments," *Journal of the American Statistical Association*, 103,1334–1344.

Shrive, F. M., Stuart, H., Quan, H., and Ghali, W. A. (2006), "Dealing with Missing Data in a Multi-question Depression Scale: A Comparison of Imputation Methods," *BMC Medical Research Methodology*, 57. DOI:doi:10.1186/1471-2288-6-57. Available at http://www.biomedcentral.com/content/pdf/1471-2288-6-57.pdf.

Skinner, B. F. (1968), *The Technology of Teaching*. New York: Appleton-Century-Crofts New York.

Wood, R. E., Kakebeeke, B., Debowski, S., and Frese, M. (2000), "The Impact of Enactive Exploration on Intrinsic Motivation, Strategy, and Performance in Electronic Search," *Applied Psychology*, 49, 263–283.

Appendix A – SPSS Certification Task A

Exercise 1

Replicate this table showing the descriptive statistics of highest year of school completed between males and females.

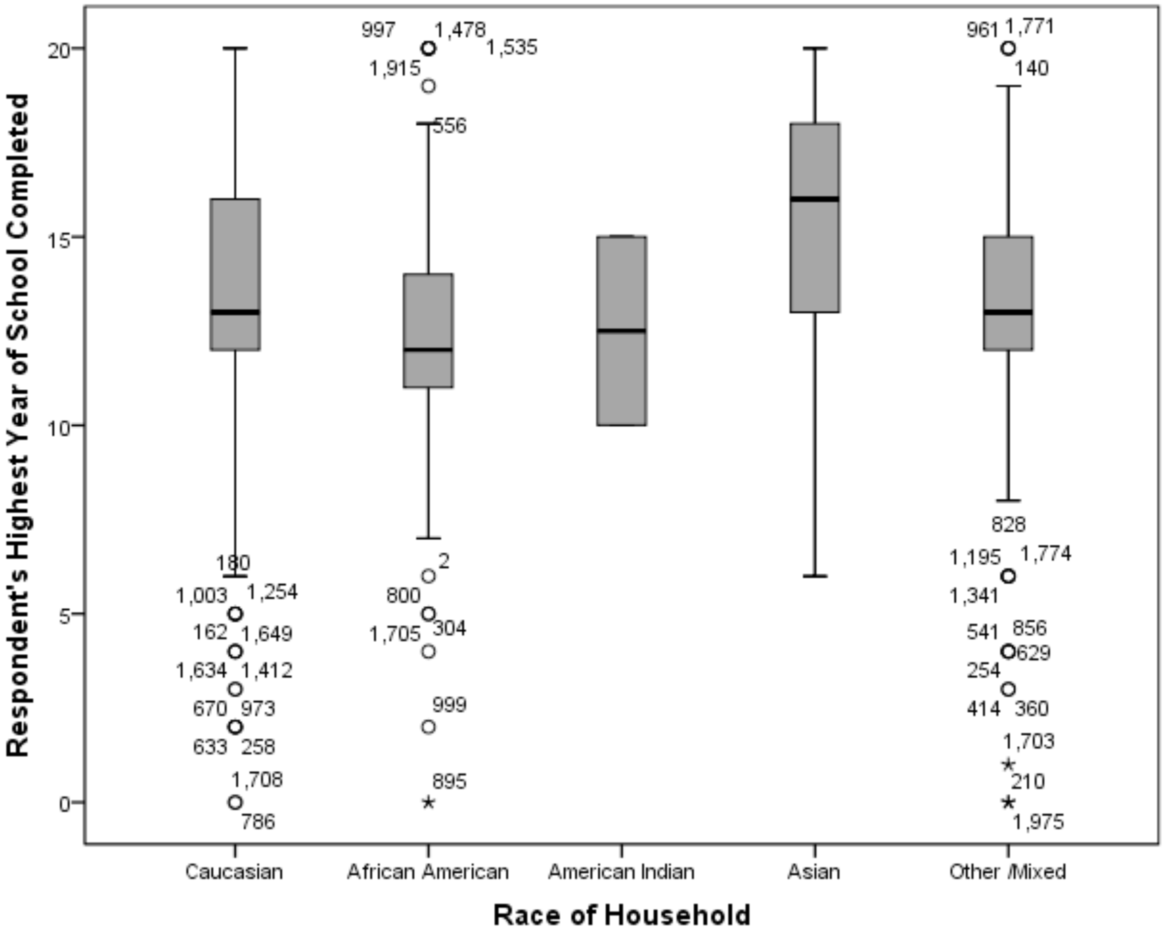
Report

Respondent's Highest Year of School Completed

Respondent's Gender	Mean	Median	N	Std. Deviation
Male	13.45	13.00	889	3.258
Female	13.47	13.00	1150	3.064
Total	13.46	13.00	2039	3.149

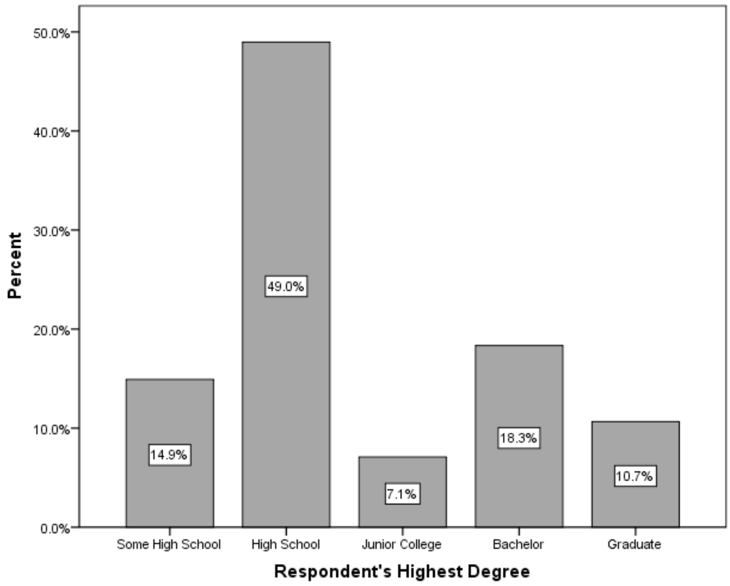
Exercise 2

Replicate this plot showing the distribution of highest year of school completed across race of household.



Exercise 3

Replicate the plot below showing the distribution of the highest level of education obtained by the sample.



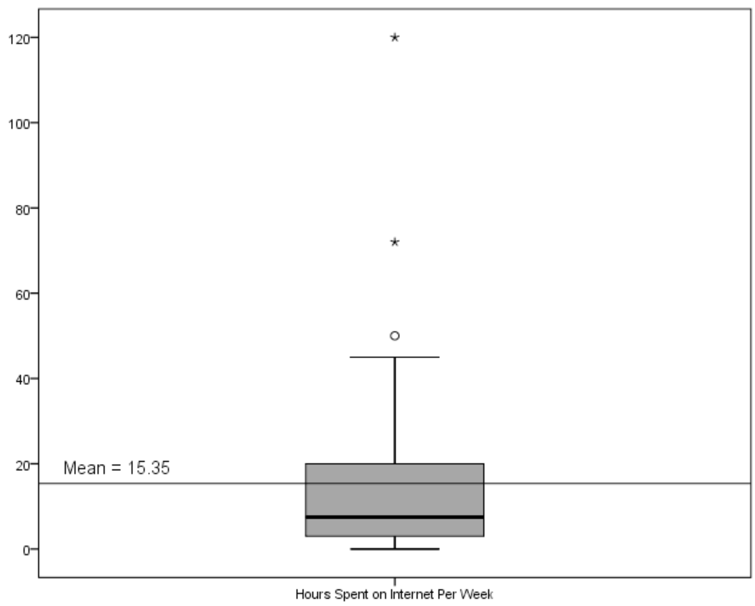
Exercise 4

Replicate the following custom table summarising the demographic characteristics of the survey sample.

			Respondent's Gender		
			Male	Female	Total
Respondent's Age	Mean		48	48	48
	SD		17	18	18
	N		890	1151	2041
Respondent's Highest Year of School Completed	Mean		13	13	13
	SD		3	3	3
	N		889	1150	2039
Estimated Family Income	Mean		50230	42015	45697
	SD		41175	37766	39532
	N		809	996	1805
Race of Household	Caucasian	Count	689	860	1549
		%	77.6%	74.8%	76.0%
	African American	Count	108	182	290
		%	12.2%	15.8%	14.2%
	American Indian	Count	0	2	2
		%	.0%	.2%	.1%
	Asian	Count	20	26	46
		%	2.3%	2.3%	2.3%
	Other /Mixed	Count	71	79	150
		%	8.0%	6.9%	7.4%

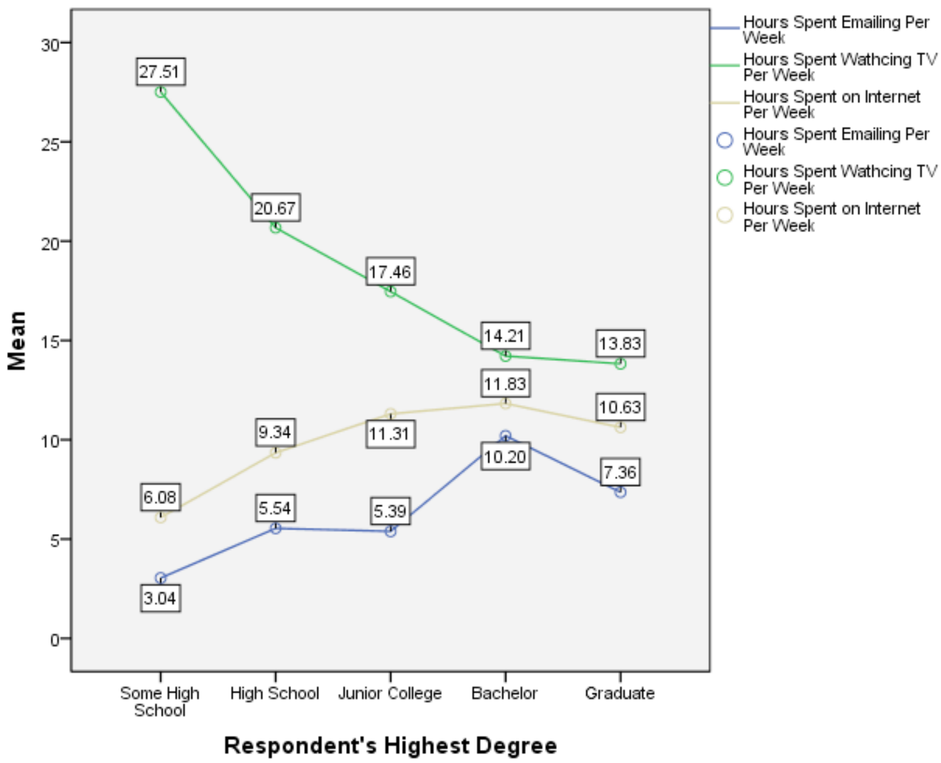
Exercise 5

Replicate the plot below showing the distribution of hours spent on the internet per week for males under the age of 25. The plot includes a reference line showing the location of the mean in comparison to the median.



Exercise 6

Replicate this plot showing the mean hours per week that respondents across different levels of education spent watching TV, using email and using the internet. The hours spent watching TV per week variable was calculated using the hours spent watching TV per day variable.



Appendix B – SPSS Certification Task A Grading Rubric

Question	Description Criteria	Marks
1a	Compare means used with correct variables – schooling and gender	/1
	Median added	/2
	Median inserted between Mean and N	/1
2a	Boxplot with correct variables – highest year of school completed by race	/2
3a	Created bar chart	/1
	Y axis shows %	/1
	Correct variables used	/1
	Value labels added	/2
4a	Age, year of schooling, family income and race of household included	/1
	Table split by gender	/2
	Total column included	/1
	Column % included for categorical variables	/2
	SD and valid N included	/1
	Mean, SD and N relabelled	/2
	Statistics positioned as rows	/1
5a	Correctly selected cases (Males < 25 years) or (Select < 25 & Split file)	/2
	Create boxplot of filtered hours spent on Internet	/1
	Add reference line for mean	/2
	Add label for reference line	/2
	Labels removed	/2
6a	Hours watching TV per week converted to hours per week	/2
	New variable labelled correctly	/1
	Line plot with highest degree on x axis	/1
	Multiple lines for each variable on one plot	/1
	Markers added	/1
	Labels added	/2
	Adaptive Transfer Total[1]:	/32

[1] Only questions 3 -6 were included for adaptive transfer scores