An Experimental Study Evaluating Error Management Training for Learning to Operate a Statistical Package in an Introductory Statistics Course: Is Less Guidance More?

James Baglin and Cliff Da Costa

Corresponding author: James Baglin (james.baglin@rmit.edu.au)
School of Mathematical and Geospatial Sciences, RMIT University, Melbourne VIC 3000, Australia

Keywords: statistics education, statistical packages, active-exploratory training, error management training, educational experiment

International Journal of Innovation in Science and Mathematics Education, *20*(3), 48-67, 2012.

Abstract

Developing the ability to operate a statistical package is a valuable student learning outcome in introductory statistics courses. Despite this, very little is known about the development of this specialised skill. This study aimed to evaluate the effectiveness of an *Error-management training* (EMT) strategy in learning to operate the statistical package *SPSS*. EMT uses minimal guidance to actively engage students in exploring the task domain and utilises errors made during training as valuable learning opportunities. EMT was compared to a conventional *Guided training* (GT) strategy which used error-avoidant, step-by-step instructions. A sample of 100 psychology students enrolled in a first year introductory statistics course were randomly allocated to either EMT or GT. Participants completed five fortnightly *SPSS* training sessions. Prior to the last training session, participants completed a post-training self-assessment task that assessed training transfer. The same self-assessment task was also completed as a follow-up in semester two. After controlling for covariates, the results of this study found no statistically significant difference between the training strategies on measures of training transfer. While a number of limitations hindered a conclusive result, issues and challenges discussed in this study provide valuable lessons for future research in this area.

Introduction

The ubiquitous nature of technology has had one of the most profound impacts on modern introductory statistics courses. Utilising technology in the statistics classroom has been proposed by proponents of statistics education reform to support student learning (Ben-Zvi, 2000). The Guidelines for Assessment and Instruction in Statistics Education (GAISE, 2005) Project report made the "use technology for developing concepts and analysing data" (p. 12) a key recommendation for improving the introductory statistics course. Survey studies confirm that the majority of recent improvements implemented in the teaching of introductory statistics courses relate to the increased use of technology (Garfield, Hogg, Schau, & Whittinghill, 2002). Statistics educational technology encompasses a wide range of tools including statistical packages (e.g. *SPSS/PASW*, *STATA*, *Minitab*, *SAS* and *R*), educational software, spread sheets, java applets, graphics calculators, multimedia, and data repositories (Chance, Ben-Zvi, Garfield, & Medina, 2007). Perhaps the most common example is the use of statistical packages.

Statistical packages are computer programs designed for the purpose of performing statistical analysis (Chance, Ben-Zvi, Garfield, & Medina, 2007). A major advantage of using statistical packages is meeting the GAISE Project report's key recommendation to focus on student's conceptual understanding of statistical topics and less on memorising the recipes, calculations and procedures of statistics (GAISE, 2005, p. 10). On a more practical note, the ability to operate a statistical package enhances students' academic careers and provides them with highly sought after workforce skills. Performing a quick search of major job websites will result in many job advertisements which specifically mention experience with the statistical packages as a key criterion for selection. A selection of such criteria include: "Proficiency with *SPSS* or similar analytics software desired", "good knowledge of *SPSS*", and "familiarity with *SPSS/PASW*". Developing these student capabilities is vital for institutions which adopt a strong work-ready focus.

Despite these advantages and the widespread adoption of statistical packages in introductory statistics courses, the development of the ability to operate a statistical package has been largely overlooked by the statistics education literature. The main reason for this oversight relates to the field's focus on the teaching and assessment of the primary educational outcomes of an introductory statistics course, namely *statistical literacy*, *reasoning* and *thinking* (see Ben-Zvi & Garfield, 2005; Garfield & Ben-Zvi, 2005). These outcomes are and should always remain the primary focus of the field. Nonetheless, the development of statistical packages skills requires some much needed attention. The literature on general software training (e.g. learning to use word processors, internet browsers, and spread sheets) provides a useful starting point.

Software training strategies can be divided into two major types, *guided* and *active-exploratory*. *Guided training* (GT) is based on the programmed learning method developed by Skinner (1968). The learner is viewed as a passive participant during training which uses step-by-step, comprehensive and explicit instructions to learn the features and procedures of a task domain (Keith, Richter, & Naumann, 2010). Mastery of the package comes through repeated practice where operational errors are avoided. In contrast, *active-exploratory* training uses minimal information to engage trainees in active exploration of a task domain (Frese, Brodbeck, Heinbokel, Mooser, Schleiffenbaum, & Thiemann, 1991). Thus, students are assumed to be active participants in the training process (Bell & Kozlowski, 2008). *Error-management training* (EMT), a well-known type of active-exploratory training, uses minimal guidance to encourage exploration and thereby increases the chances of making errors. According to EMT, errors are argued to be beneficial to training as they promote deeper exploration, help develop the know-how to avoid errors and the know-how to overcome errors once they have been committed (Frese et al., 1991). EMT frames errors in a positive light by presenting heuristics to students during training such as "Errors are a natural part of learning. They point out what you can still learn!'" (Dormann & Frese, 1994, p. 368) .

As Keith et al. (2010) proposes, EMT works by developing a trainee's self-regulatory skills, *metacognition* and *emotional control*, more effectively than GT. Metacognition, defined as the ability to exert "control over his or her cognitions" (Ford, Smith, Weissbein, Gully, & Salas, 1998, p. 220), is developed by EMT through exploration. Exploration require students to practice the three basic processes of metacognition - planning, monitoring, and evaluating (Brown, Bransford, Ferrara, & Campione, 1983). GT, on the other hand, largely ignores these processes above and beyond what is required to follow step-by-step instructions. EMT is also argued to help trainees develop their emotion control which is 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, Ackerman, & Heggestad, 1996, p. 186). EMT is argued to achieve this by framing errors in a positive light. GT is error-avoidant and pays no particular attention to the development of emotion control.

The effectiveness of both EMT and GT has been evaluated using measures of training transfer. Training transfer can be defined as knowledge and skills gained during training which transfer to other tasks and jobs outside of training (Hesketh, 1997). Keith and Frese (2008) differentiated between two major types of training transfer, *analogical* and *adaptive*. *Analogical transfer* relates to tasks that are similar to those covered during training (Keith & Frese, 2008; Keith et al., 2010), whilst on the other hand, *adaptive transfer* tasks consists of tasks that are structurally distinct from training and require the trainee to adapt their knowledge gained from training in novel ways (Ivancic & Hesketh, 1996; Keith et al., 2010). Given the acute nature of training in university settings, adaptive transfer is considered a more desirable capability as it promotes sustainable learning outside of training. Training strategies which promote adaptive transfer are more advantageous to students.

A meta-analysis by Keith and Frese (2008) found EMT for general software training to be superior to GT on transfer performance. Keith and Frese analysed 24 studies comparing EMT to GT for a wide variety of software including simulation, word processing, databases, presentations, spreadsheets, e-mail, web browsers, programming languages and statistical packages. The meta-analytic results of combining these studies found that EMT was moderately superior to GT on measures of analogical transfer and substantially more effective for adaptive transfer. Keith et al. (2010) explain that EMT is particularly effective for adaptive transfer because self-regulatory skills are vital to adaptive transfer tasks and these skills are better developed in EMT than GT. In addition, Keith and Frese also found that the two elements of EMT (exploration and error encouragement) contributed unique training benefits. This suggests that active-exploratory training alone can be enhanced with the explicit encouragement of errors. One published study from this meta-analysis looked at statistical package training.

An experiment by Dormann and Frese (1994) randomly allocated 30 psychology students to either EMT or GT training to use the statistical package *SPSS*. While Dormann and Frese did not differentiate between analogical and adaptive transfer, the results of the study found that the EMT group out-performed the GT group on both moderate and difficult transfer tasks. However, the Dormann and Frese study had a number of limitations. The experiment used a small sample, evaluated transfer immediately after training proving no useful measure of real-world retention (e.g. week to week, semester to semester), used only a single training session outside of a regular course and used a very early version of the statistical package *SPSS* that differs substantially from current day versions. While the results were promising, there is a clear need for current research evaluating EMT in real introductory statistics courses, using larger samples and assessing training transfer at more meaningful follow-up. Other training outcomes should also be considered. Students' experiences and perceptions of training will impact on instructors' decisions to use particular training strategies in their courses.

Consequently, the aim of this study was to evaluate the effectiveness of EMT versus GT for learning to use a statistical package over the duration of a one-semester introductory statistics course. Preliminary work was presented at the 7th Australian Conference on Science and Mathematics Education (Baglin, Da Costa, Ovens & Bablas, 2011). A qualitative arm to this project has also been reported in Baglin and Da Costa (2012). In line with previous research,

it was hypothesised that EMT would be comparable to GT for analogical transfer tasks, but that EMT would be superior to GT for adaptive transfer tasks. A second aim was to evaluate possible advantages and disadvantages of implementing either strategy. This study considered students' perceptions of training satisfaction, training difficulty, statistical package self-efficacy and training anxiety.

Method

Participants

Participants consisted of 1st year psychology students enrolled in an introductory statistics course which ran concurrently across two campuses. Students were randomly assigned to odd and even week computer laboratory sessions as part of a regular course requirement. Of the 151 students enrolled, 117 consented to participate in the experiment. Three of these consenting students were not randomly allocated but instead placed automatically into available labratories due to space limitations. There were 14 consenting participants who did not finish training. Seventy-six of these consenting students who finished training completed a post-training follow-up questionnaire. Seventy-nine of the consenting students that finished training in semester one were followed-up in a semester two statistics course. A flowchart summarising the study is shown in Figure 1. Table 1 displays the characteristics of the sample across the EMT and GT strategies.

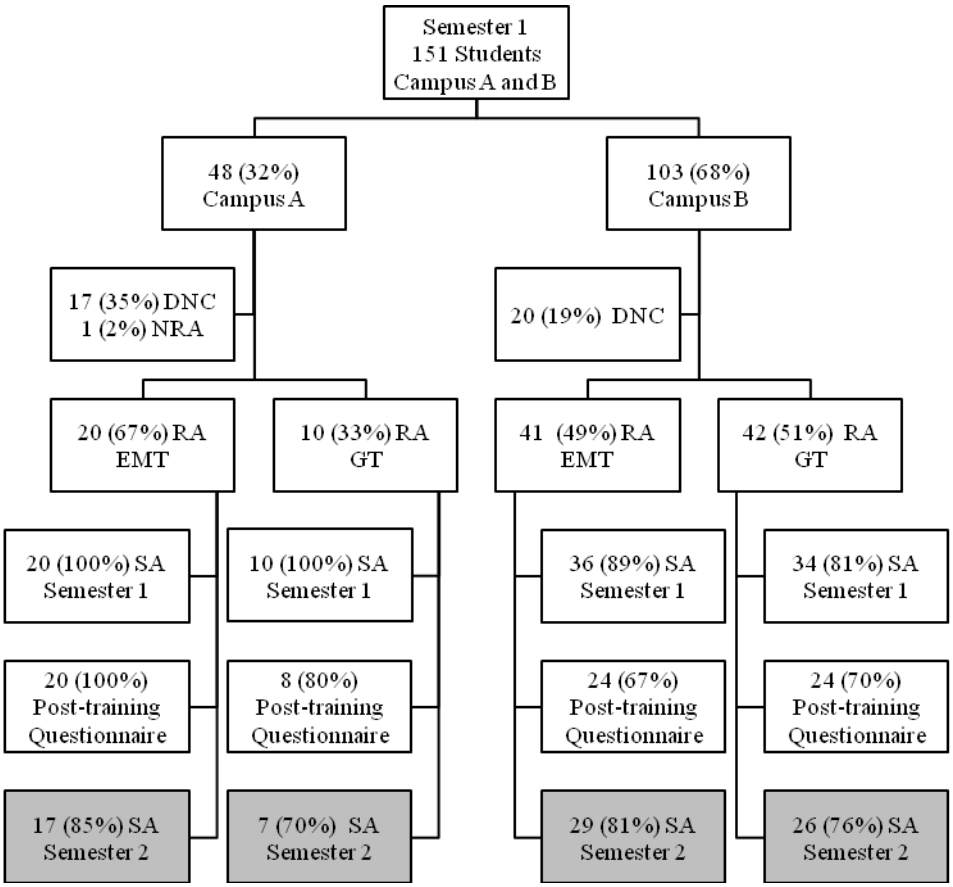


Figure 1: Study flow chart. *Note*. RA = Randomly allocated, NRA = Not randomly allocated, DNC = Did not consent, EMT = Error-management training, GT = Guided Training, SA = Completed self-assessment 1 & 2. Semester 2 follow-up has been shaded.

Measures

Covariates

Statistical knowledge, which was defined as the proportion of marks obtained on the end of semester multiple-choice exam, was included as a covariate in the statistical analysis of the results. Statistical knowledge scores were used to control for the influence of statistical knowledge on operating the statistical package. Even though this study employed random allocation to training strategies to help reduce group bias, controlling this covariate would facilitate a more accurate comparison of the two training strategies.

Table 1: Sample Characteristics of Strategies

		Strategy		
		GT	EMT	Total
Strategy	*N* (%)	44 (44%)	56 (56.0%)	100
Campus A	*N* (%)	10 (33.3%)	20 (66.7%)	30
Campus B	*N* (%)	34 (48.6%)	36 (51.4%)	70
Female	*N* (%)	31 (44.3%)	39 (55.7%)	70
Male	*N* (%)	13 (43.3%)	17 (56.7%)	30
Age	*M* ± *SD*	19.84 ± 5.02	19.41 ± 5.07	19.60 ± 5.05

Note. Table adapted from Baglin et al. (2011).

Training adherence was monitored throughout the semester in order to take into account the extent to which a participant engaged in training. Adherence was measured by two indicators - laboratory completion and laboratory compliance. Completion was defined as finishing a laboratory training session, whereas compliance was defined as attending an allocated laboratory training session. To construct this score, the number of completed training laboratory sessions was added to the number of times a participant completed their training laboratory sessions during their designated times. If they completed any laboratory session in a different week or during their own time, compliance was scored as zero for that laboratory session. Due to a system error with logging laboratory session 1 grades, only laboratory sessions 2 - 5 were included for the calculation of this score. Therefore, the training adherence scores could range from no adherence (0) to perfect adherence (8).

Self-assessment compliance was also taken into consideration. Compliance was defined as whether the student completed both self-assessment tasks of training transfer in the allocated self-assessment laboratory session. If the students completed any of the self-assessment tasks outside of the allocated self-assessment laboratory session, they were classified as non-compliant. Compliance was important to take into account as students who did not attend the scheduled self-assessment laboratory sessions were not under supervision. These non-compliant students could have gone over the allocated time limit or received assistance from peers who had already completed the self-assessment tasks. Therefore, non-compliance was hypothesised to be associated with inflated self-assessment scores and would need to be controlled for when comparing training strategies on training transfer.

Training Outcomes

Measures of training transfer were obtained using two self-assessment tasks that were completed in the final weeks of training between laboratory sessions 4 and 5. The same self-assessment tasks were also used in the semester two follow-up. Self-assessment 1 consisted of eight exercises that measured a student's analogical transfer. These exercises were similar

to tasks completed during training. Self-assessment 2 consisted of eight exercises that measured adaptive transfer. Adaptive transfer tasks were structurally distinct from training and required students to complete tasks and analyses in SPSS that were not strictly covered during training. This included completing highly difficult tasks, novel tasks that were similar but not explicitly covered, and linking multiple tasks together in novel ways. Only four of these exercises at post-training were included due to technical difficulties with the online self-assessment. All eight adaptive items were included at follow-up. A total transfer score was also computed by summing analogical and adaptive transfer scores.

Students were given 25 minutes to complete each self-assessment task. However, as students were able to complete laboratory sessions outside of allocated laboratory session times, this should be considered a soft time limit. Students were instructed that to obtain a grade for the self-assessment, they would need to obtain at least 4/8 on self-assessment 1 and 2/4 on self-assessment 2. Questions were randomised from pools of similar questions. Participants were allowed to attempt each self-assessment up to 5 times as the laboratory sessions and self-assessment were graded on completion (formative assessment). For evaluating the effect of training strategies, only a participant's first attempt on each self-assessment was used.

When designing the self-assessment tasks, it was important that each task measured a student's ability to successfully operate the statistical package and not be confounded by the student's knowledge of statistics. For example, completing an exercise task that requires a student to find the median IQ of the sample may be confounded by the student's knowledge of the median. Each exercise was designed to minimise this dependency. For example, exercise questions which were used to score someone on their ability to operate SPSS asked questions relating to the acquired output from SPSS that proved they had completed the analysis correctly. The questions avoided interpretation of statistics or graphs which would be dependent on student's statistical knowledge. While it would be impossible to completely remove this dependency, the inclusion of a statistical knowledge covariate would help to further control this dependency when comparing strategies.

A post-training questionnaire also asked students to rate their perceptions of training difficulty, training satisfaction, training anxiety, and statistical package self-efficacy. These measures were included to consider other important outcomes of training that might be of concern to instructors. The training difficulty and satisfaction items were rated on a seven-point likert-type scale where (1) referred to very easy/not at all satisfied and (7) referred to very difficult/completely satisfied respectively.

Anxiety during statistical package training was measured using four items adapted from the Tension-pressure dimension scale of the Intrinsic Motivation Inventory by Deci and Ryan reported in McAuley, Duncan, and Tammen (1989). A sample item that was adapted is "I felt pressured when training to use SPSS". These items were responded to on a seven-point likert-type scale ranging from strongly disagree (1) to strongly agree (7). Ratings on each of these four items were average to get an overall training anxiety rating score where higher scores are indicative of higher training anxiety. The results of a principle components analysis (PCA), using an eigen value greater than one criteria for component selection, resulted in a single component which explained 56.01% of the variation in training anxiety scores. Internal consistency of the scale resulted in Cronbach's $\alpha = .74$.

Statistical package self-efficacy, defined as a participant's confidence in their ability to operate a statistical package after training, was measured using three items from Finney and

Schraw's (2003) Current Statistics Self-efficacy (CSSE) scale. 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 $\alpha = 0.78$). A PCA extracted a single construct which explained 74.23% of the variation in responses.

Manipulation Checks

Manipulation checks were measured across both strategies using items contained in the self-reported post-training questionnaire. All items were responded to on a seven-point likert-type scale ranging from strongly disagree (1) to strongly agree (7). All items were borrowed or adapted from previous research. Scales composed of multiple items were averaged to get a final scale score. The manipulation checks were used to validate the correct implementation of the training strategies. It was hypothesised that the EMT strategy would be associated with higher self-reported metacognitive activity, emotional control, error-orientation, and exploration.

The degree to with students engaged in metacognitive activity during training was measured using 12 items from a self-report scale heavily adapted from Ford et al. (1998). The items asked questions relating to the extent to which a participant engaged in metacognitive activities during training (i.e. monitoring, planning and revising). A sample item is "When my methods were not successful for completing statistical procedures in SPSS, I experimented with different approaches for completing the procedure". Higher scores indicate a higher self-reported level of metacognitive activity during training. Due to the substantial adaptation of the original Ford et al. items, the psychometric properties of the scale items were re-checked. A PCA extracted a single component using the eigen value greater than one approach which explained 50.54% of the variability in responses to metacognitive activity items. Cronbach's α for the adapted scale was .91.

The degree to which students exhibited emotional control during training was checked using eight items adapted from Keith and Frese (2005). These items related to the degree to which participants regulated their emotions during training. An example of an item is "When difficulties arose during computer labs I did not allow myself to lose my composure". According to a PCA of the adapted items, a unidimensional component explained 55.3% of the variation in responses to the emotional control items. The emotional control scale had high internal consistency with Cronbach's $\alpha = 0.89$.

Error-orientation, or a participant's attitude towards errors made during training, was measured using two subscales adapted for statistical package training from the Error Orientation Questionnaire (EOQ, Rybowiak, Garst, Frese, & Batinic, 1999). The original EOQ was developed to measure how employees cope with errors committed in the workplace. The two subscales of EOQ, Error Strain (five items, e.g. "When I made a mistake in SPSS, I lost my temper and got angry about it") and Learning from Errors (four items, e.g. "From my errors, I have learned a lot about how to work with SPSS") had high internal consistency with $\alpha = .79$ and .89 respectively (Rybowiak et al., 1999). These original items were adapted to refer specifically to using SPSS. High scores for Learning from Errors indicate a positive attitude towards errors and high scores on Error Strain indicate an emotional intolerance for errors. A PCA confirmed the two-dimensional structure of the EOQ

with Learning from Errors accounting for 35.76% and Error Strain accounting for 28.26% of the variability in responses. Cronbach's α for learning and error strain was .86 and .80 respectively.

The extent to which participants engaged in exploratory behaviour versus guided instruction during training was measured using six items based on Bell and Kozlowski (2008). Three items which related to GT included the use of step-by-step instructions (e.g. "I used step-by-step instructions when learning to use SPSS"), copying other students (e.g. "I copied how other students completed tasks in SPSS."), and seeking assistance from tutors ("When I was unsure about how to complete a task in SPSS, I would immediately ask the tutor/or a friend for help"). Another three items related to EMT (e.g. "I explored the features of SPSS without much instruction by changing options or trying different analyses in order to complete each laboratory exercise").

A PCA on these six items revealed two components (Eigen values greater than one). The first component, labelled "Active" explained 35.85% of the variation in responses, whereas the second component, labelled "Guided" explained 20.76%. Cronbach's α was .69 and .41 for Active and Guided components respectively. Due to the unimpressive coefficients and the fact that these items appeared to assess somewhat unrelated aspects of guided and active-exploratory training, it was decided to individually compare each item's mean self-reported rating between strategies when checking the validity of manipulations.

Procedure

Following university ethics approval and random allocation to odd and even week computer laboratory sessions, students were approached before their first lecture to participate in the study. Non-consenting students were still required to complete training, but their data was not collected for the purpose of this study. The allocation to odd and even week laboratory sessions was due to restrictions with computer laboratory availability. This odd and even week group allocation allowed for the manipulation of training strategies. The ordering of EMT and GT to odd and even weeks was counterbalanced between the campuses. Campus A had GT on odd weeks and EMT training on even weeks. On campus B the order was reversed.

Training consisted of five laboratory sessions for training to use the statistical package SPSS. Topics included the following: 1) SPSS Introduction, 2) SPSS Basics, 3) Frequencies and Bar Charts, 4) Cross-tabs and Chi-square tests and 5) Correlation and Regression. Specifically, this study used version 18 which was temporarily re-named for legal reasons to PASW 18 in 2009. Since being acquired by IBM® in 2010, the package has been re-named to IBM SPSS. To avoid confusion, SPSS will be used throughout this article even though students used PASW 18.

Self-assessment tasks measuring training transfer outcomes were completed towards the end of the semester between laboratory sessions 4 and 5. The same self-assessment tasks were repeated again two months after the completion of training for follow-up in the first two weeks of semester two. Laboratory sessions were scheduled for one hour per week. However, students were permitted to stay longer to finish or catch-up. Students who missed their designated laboratory sessions were required to ask permission to attend a non-designated laboratory session. This was done so as to not disadvantage students and was a condition for ethics approval. This meant that some students mixed between strategies and could not be

blinded. Blinding was also limited by the fact that participants would have talked to each other. However, the exact nature of the strategies was withheld from students until the completion of training.

Training was delivered using a streamlined, proprietary, online assessment system called WebLearn. WebLearn is similar to a streamlined version of Blackboard's quiz, test and assignment features. Each laboratory session consisted of objectives, instructions and exercises embedded with the strategy's instructions. Students would sequentially work through each exercise which was designed to introduce them to and get them practising the SPSS features related to the course content. To show that the student had successfully completed the procedure in SPSS, each exercise required students to answer a question that could only be answered if they had correctly operated SPSS. Students were required to score 70% or above to gain a participation mark. Students were allowed to reattempt laboratory sessions. To find out if they had passed the laboratory session, the student would submit all their answers to the WebLearn system for grading when they had completed all the laboratory session exercises. Marking was done automatically by WebLearn.

The GT group received comprehensive step-by-step instructions and screen shots summarising each exercise in SPSS (Figure 2a). These students were instructed to follow these steps and try to avoid making errors. The EMT strategy was given the exact same exercises but with modified instructions and no screen shots. The EMT training used minimal guidance to get the participant actively exploring SPSS (Figure 2b). Instructions were designed to point the students in the right direction, but students were left to work out the specifics. Sometimes for difficult analyses, hints were given to help students get back on track if they veered too far from the correct path. Students were also presented with error management heuristics listed at the top of each exercise. Examples of these heuristics included "If you have a problem, regard it as a learning opportunity" (Wood, Kakebeeke, Debowski, & Frese, 2000) and "Errors are a natural part of learning. They point out what you can still learn". These heuristics aimed to frame errors in a positive way to help students develop emotion control and benefit from the insight that can be gained through their errors.

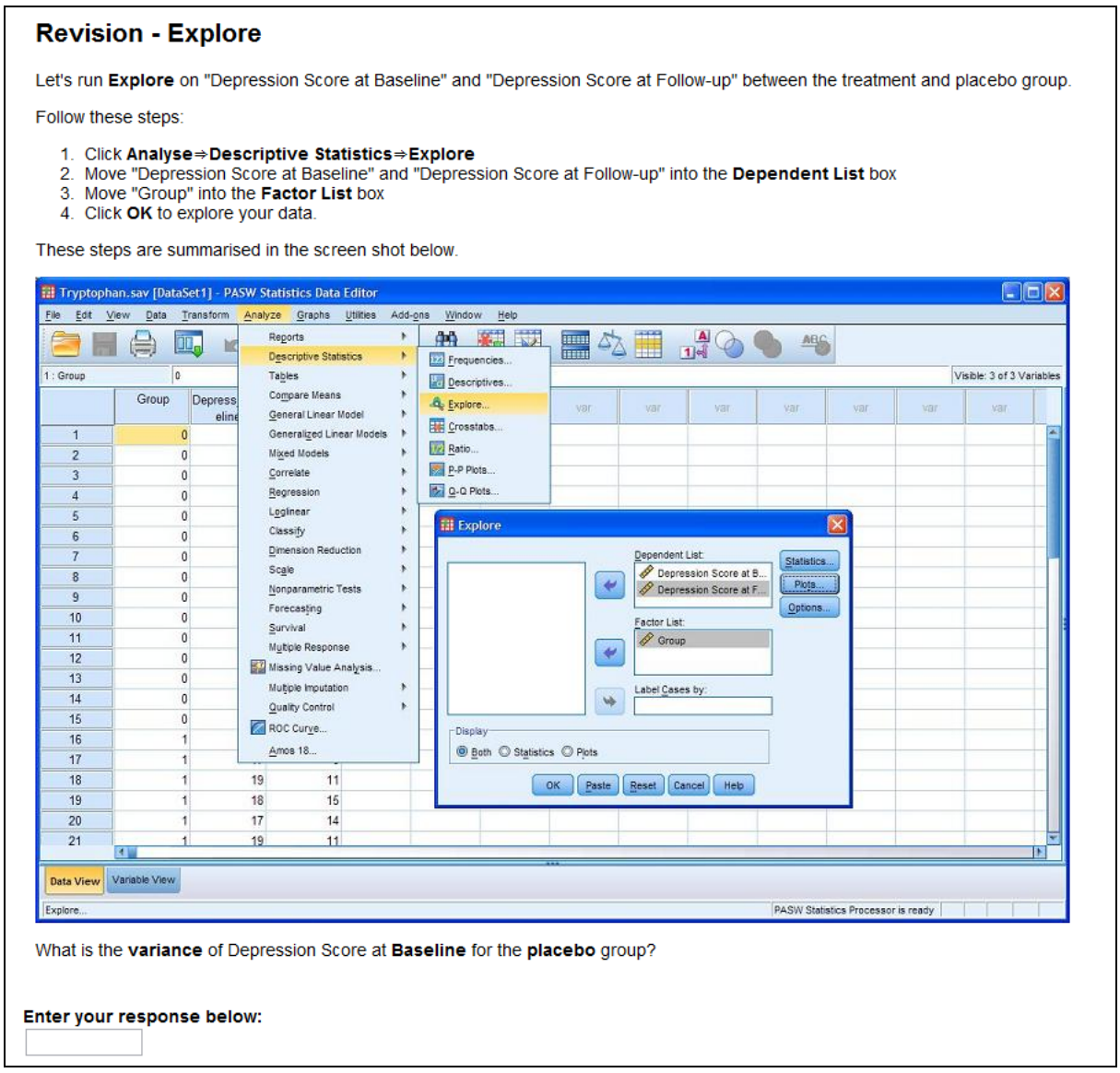
A laboratory supervisor was also present at each scheduled laboratory session. In the GT strategy, the supervisor was instructed to help the students as much as they needed in line with the theory of GT. In the EMT strategy, the supervisor was advised to encourage the students to find the solution themselves. If the participant was struggling after multiple attempts, the supervisor was allowed to give them a hint to get them back on track. The supervisor was also trained to reinforce the positive error framing by encouraging students to learn from their mistakes.

In the final lecture following semester one's training, students were approached to fill out the self-reported post-training questionnaire which contained the manipulation check and other training outcome items (difficulty, satisfaction, self-efficacy and anxiety).

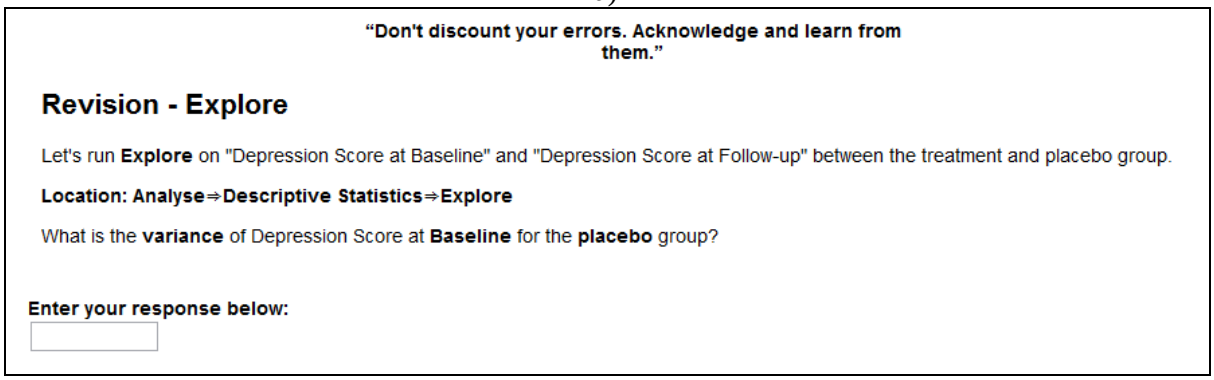
Results

Descriptive statistics and intercorrelations for training transfer outcomes and covariates are shown in Table 2. For the covariates, the EMT strategy had higher mean training adherence and post-training compliance, but lower statistical knowledge and follow-up training compliance when compared to the GT strategy. Descriptively at post-training, the EMT strategy outscored the GT strategy on analogical and total training transfer scores, but not on

adaptive transfer. At follow-up, the EMT group out-scored the GT group on adaptive transfer, but the GT strategy appeared to do better on analogical and total transfer scores. The next stage was to statistically model these training outcomes after controlling for covariates.



a)



b)

Figure 2: a) An example of *GT* instructions for learning to explore variables in *SPSS*. b) An example of an *EMT* exercise for learning to explore variables in *SPSS*.

Table 2: Descriptive Statistics and Intercorrelations for Training Transfer Measures and Covariates

Variable		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. Training Adherence		-	.270**	.137	.146	.174	.122	.061	.104	.379**	.168
2. Statistical Knowledge			-	.299**	.219*	.325**	.451**	.431**	.507**	.124	.043
3. Analogical Transfer 1				-	.321**	.917**	.433**	.369**	.460**	-.176	-.177
4. Adaptive Transfer 1					-	.663**	.214	.412**	.364**	-.181	-.119
5. Total Transfer 1						-	.426**	.459**	.509**	-.205*	-.195
6. Analogical Transfer 2							-	.513**	.861**	-.097	-.024
7. Adaptive Transfer 2								-	.878**	.006	-.059
8. Total Transfer 2									-	-.050	-.049
9. SA Compliance 1										-	-.042
10. SA Compliance 2											-
EMT	*M*	7.05	.69	5.41	1.66	7.07	7.54	2.89	10.43	71.4%	84.8%
	SD	1.38	.15	1.66	.94	2.16	1.53	1.90	2.93		
	N	56	56	56	56	56	46	46	46	56	46
GT	*M*	6.84	.74	5.23	1.72	6.91	7.75	2.69	10.44	47.7%	96.9%
	SD	1.40	.15	1.92	.88	2.31	1.97	1.73	3.33		
	N	44	41	44	43	44	32	32	32	44	32
Total	*M*	6.96	.71	5.33	1.69	7.00	7.63	2.81	10.44	61.0%	89.7%
	SD	1.38	.15	1.77	.91	2.22	1.71	1.82	3.08		
	N	100	97	100	99	100	78	78	78	100	78

Note. SA = Self-assessment, 1 = Post-training (1st Semester), 2 = Follow-up (2nd semester).
* $p < .05$, ** $p < .01$

One-way analysis of covariance (ANCOVA) was performed to assess for significant differences between the GT and EMT strategies on mean post-training and follow-up transfer outcomes (see Table 3). The ANCOVA used training adherence, self-assessment compliance and statistical knowledge as covariates. Table 3 contains the ANCOVA model parameters and covariate adjusted means with 95% *CI* for all three training transfer outcomes across post-training and follow-up. The partial η^2 statistic has been included as a estimate of effect size. The η^2 statistic reflects the proportion of variability in an outcome variable that can be explained by its relationship with a particular variable after controlling for the effects of other variables in a model.

The primary focus of the ANCOVA models was to compare the strategies on training transfer outcomes after controlling for statistical knowledge, training adherence, and self-assessment compliance (Table 3). According to the post-training outcomes there were no statistically significant differences between strategies on mean analogical, $F(1,92) = 2.25$, $p = 0.137$, $\eta^2 = .02$, adaptive, $F(1,91) = 0.10$, $p = .754$, $\eta^2 = .00$ and total training transfer scores, $F(1,92) = 2.08$, $p = 0.153$, $\eta^2 = .02$, after controlling for covariates (Figure 3). The same non-significant trend was found at follow-up, analogical, $F(1,73) = 0.001$, $p = 0.978$, $\eta^2 = 0$, adaptive, $F(1,73) = 1.47$, $p = .23$, $\eta^2 = .02$ and total training transfer scores. $F(1,73) = 0.59$, $p = 0.447$, $\eta^2 = .008$ (Figure 3).

In all models, except for adaptive transfer at post-training, statistical knowledge was a statistically significant positive covariate (Table 3). This indicated that there was a positive relationship between training transfer outcomes and statistical knowledge. In addition to this finding, at follow-up in semester two the effect of statistical knowledge increased (see η^2 in Table 3). This suggests that as the gap between training completion and follow-up increases, the ability to operate a statistical package becomes more dependent on a student's knowledge of statistics. Compliance was also a statistically significant covariate for all outcomes at post-training, but not for follow-up. According to the ANCOVA models in Table 3, compliance was associated with lower transfer scores. This supported the belief that non-compliers were at a significant advantage on self-assessment tasks when compared to participants that completed self-assessment tasks under controlled conditions. The effect of self-assessment compliance at follow-up was probably less pronounced as overall compliance at follow-up was much higher.

As the results of the ANCOVA models failed to find any statistically significant differences between strategies on mean training transfer outcome scores, it was important to evaluate the validity of the imposed training strategies. A series of independent sample *t*-tests were performed comparing the responses to the self-reported manipulation check items responded to on the post-training questionnaire (Table 4 and Figure 4). The results of these comparisons found only one statistically significant difference between strategies on the self-reported use of step-by-step instructions. The EMT strategy rated a mean level of agreement on this item statistically significantly lower than the GT strategy. Surprisingly, none of the other nine comparisons were statistically significant (Table 4).

The second aim of the study was to consider other practical outcomes of using different training strategies. Students' mean ratings of training difficulty, satisfaction, anxiety and statistical package self-efficacy were compared using independent sample *t*-tests. Table 5 displays the means between the strategies and the results of the four independent sample *t*-tests. The results of this analysis found no statistically significant difference between strategies on student self-reports (Table 5).

Table 3: ANCOVA Models Predicting Training Transfer Measures

	Analogical			Adaptive			Total Transfer		
Parameters	*B*	95% *CI*	η^2	*B*	95% *CI*	η^2	*B*	95% *CI*	η^2
Post Training (Semester 1)									
Statistical Knowledge	2.42	(1.47, 6.13)	0.10	1.26	(-0.01, 2.52)	0.04	4.97	(2.12, 7.82)	0.12
Training Adherence	0.16	(-0.1, 0.42)	0.02	0.13	(-0.01, 0.27)	0.03	0.30	(-0.02, 0.62)	0.04
SA Compliance 1[a]	-1.08	(-1.82, -0.34)	0.08	-0.50	(-0.91, -0.10)	0.06	-1.54	(-2.44, -0.64)	0.11
Strategy[b]	-0.52	(-1.21, 0.17)	0.02	-0.06	(-0.43, 0.32)	0.00	-0.61	(-1.45, 0.23)	0.02
GT Adjusted Mean	5.06	(4.55, 5.57)		1.66	(1.38, 1.94)		6.69	(6.06, 7.32)	
EMT Adjusted Mean	5.58	(5.15, 6.02)		1.72	(1.49, 1.96)		7.30	(6.77, 7.83)	
Follow-up (Semester 2)									
Statistical Knowledge	5.83	(3.01, 8.66)	0.19	6.45	(3.44, 9.46)	0.20	12.28	(7.41, 17.16)	0.26
Training Adherence	0.09	(-0.23, 0.41)	0.00	-0.04	(-0.38, 0.30)	0.00	0.05	(-0.50, 0.60)	0.00
SA Compliance 2[a]	-0.30	(-1.51, 0.92)	0.00	-0.29	(-1.58, 1.00)	0.00	-0.59	(-2.68, 1.51)	0.00
Strategy[b]	-0.01	(-0.77, 0.75)	0.00	-0.50	(-1.31, 0.32)	0.02	-0.51	(-1.82, 0.81)	0.01
GT Adjusted Mean	7.62	(7.05, 8.19)		2.52	(1.91, 3.13)		10.14	(9.15, 11.13)	
EMT Adjusted Mean	7.63	(7.16, 8.11)		3.01	(2.51, 3.51)		10.64	(9.83, 11.46)	

[a] Compliant students = 1 [b] GT = 1 [c] Means after adjusting for the covariates of training adherence, SA compliance and statistical knowledge. 1 = Post-training (1st Semester), 2 = Follow-up (2nd semester).

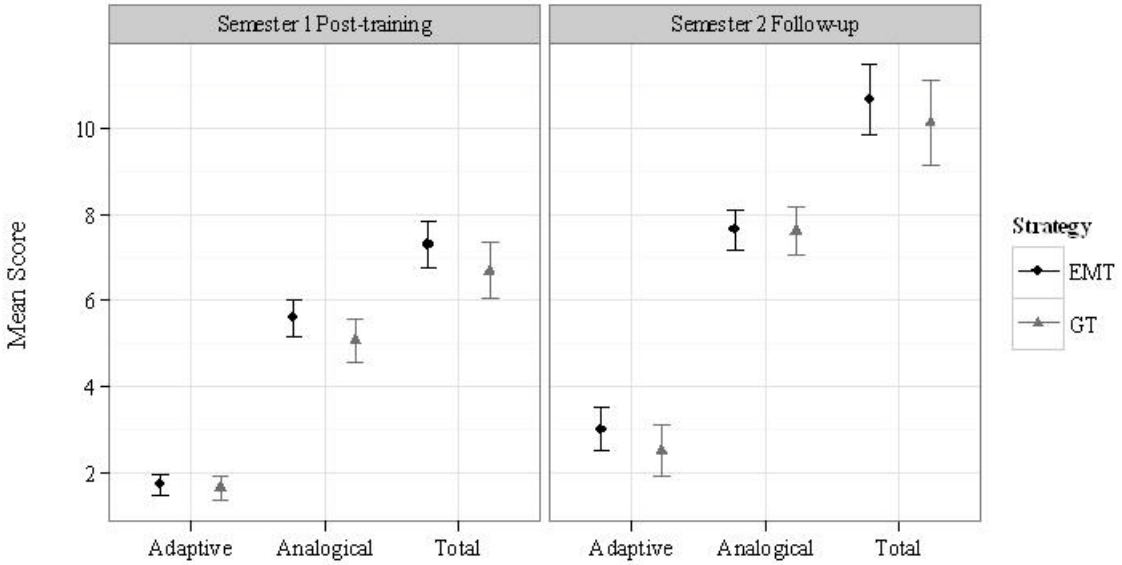


Figure 3: Training transfer covariate adjusted outcome means across strategies. Error bars show 95% *CI* of adjusted means. *Note.* Post-training adaptive transfer was scored out of 4 and follow-up adaptive transfer was scored out of 8.

Table 4: Descriptive Statistics and Independent Sample *t*-tests Comparing Mean Training Strategy Manipulation Check Scales and Items

								95% CI of Difference	
Manipulation Variable		*M*	*SD*	*N*	*SEM*	*t*	*p*	Lower	Upper
Metacognition	GT	4.06	1.03	33	.18	-0.98	.33	-.69	.24
	EMT	4.29	1.01	45	.15				
Emotional Control	GT	3.93	.60	33	.10	-0.70	.49	-.34	.17
	EMT	4.02	.52	45	.08				
Learning from Errors	GT	4.01	1.26	33	.22	-1.65	.10	-1.05	.10
	EMT	4.48	1.25	45	.19				
Error Strain	GT	3.47	1.17	33	.20	-0.47	.64	-.76	.47
	EMT	3.62	1.47	45	.22				
Used step-by-step instructions	GT	6.58	.66	33	.12	3.23[1]	<.001	.33	1.40
	EMT	5.71	1.62	45	.24				
Copied other students	GT	3.52	2.06	33	.36	1.22	.23	-.35	1.48
	EMT	2.95	1.94	44	.29				
Immediately sought assistance	GT	5.15	1.62	33	.28	1.64	.11	-.14	1.47
	EMT	4.49	1.87	45	.28				
Explored without instruction	GT	3.39	2.06	33	.36	-1.34	.18	-1.40	.27
	EMT	3.96	1.64	45	.24				
Operate without instruction	GT	3.61	1.98	33	.35	-1.10	.28	-1.36	.39
	EMT	4.09	1.87	45	.28				
Actively explored *SPSS*	GT	3.91	1.79	33	.31	-0.12	.90	-.80	.70
	EMT	3.95	1.51	44	.23				

[1] Equal variance not assumed.

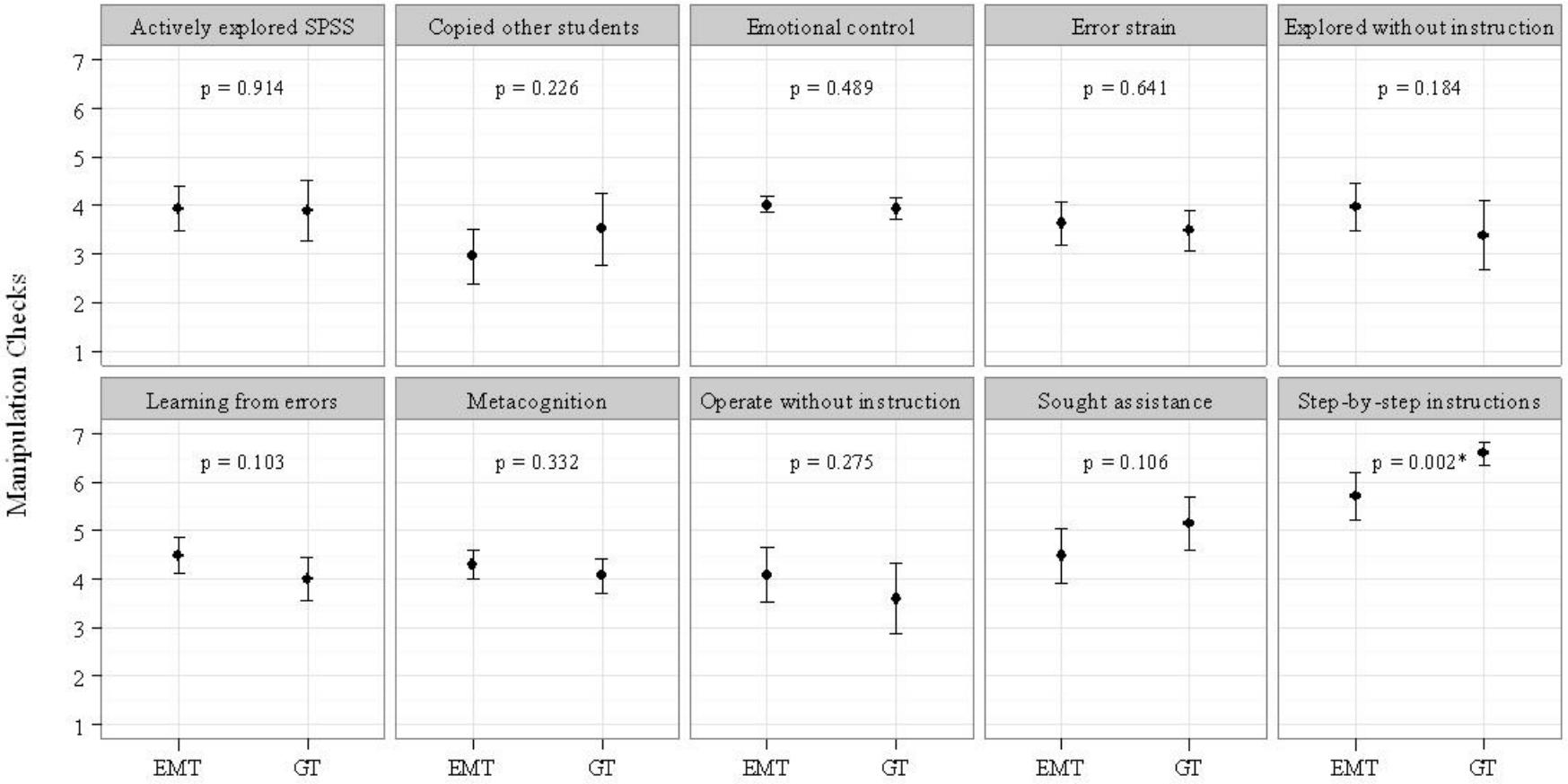


Figure 4: Error-bar plots showing 95% *CI* of the mean scores of the self-reported manipulation check items between strategies.

Table 5: Descriptive Statistics and Independent-sample *t*-tests Comparing Mean Training Difficulty, Satisfaction, Statistical Package Self-efficacy and Training Anxiety between Strategies

								95% CI of Difference	
Outcome		*M*	*SD*	*N*	*SEM*	*t*	*p*	Lower	Upper
Training Difficulty	GT	3.85	1.58	33	.28	-1.63	.11	-1.18	0.12
	EMT	4.38	1.28	45	.19				
Training Satisfaction	GT	4.24	1.66	33	.29	-0.51	.61	-0.88	0.52
	EMT	4.42	1.44	45	.21				
Self-efficacy	GT	4.64	1.13	33	.20	0.16	.88	-0.51	0.60
	EMT	4.60	1.26	45	.19				
Anxiety	GT	4.03	1.23	33	.22	-1.41	.16	-0.96	0.17
	EMT	4.43	1.22	45	.18				

Discussion

The results of this study found no statistically significant difference between EMT and GT strategies on measures of analogical, adaptive, and total training transfer at both post-training and follow-up after controlling for statistical knowledge, training adherence and self-assessment compliance. These findings failed to support the hypothesis of this study and failed to support the findings of previous research (Keith & Frese, 2008; Keith et al., 2010, Dormann & Frese, 1994) .

Statistical knowledge was the only reliable and significant predictor of training transfer performance. This study also showed that this dependency became stronger with time between post-training and two-month follow-up. There are two likely interpretations for this finding. The first suggests that a student's ongoing ability to operate a statistical package is largely dependent on their knowledge of statistics. However, an alternate interpretation is that the self-assessment tasks were largely measuring statistical knowledge instead of the ability to operate a statistical package. This study assumed that after controlling for statistical knowledge, the remaining variability in transfer scores could be attributed to variability in statistical package skills. However, there is no direct way to test this assertion. Further research is needed to better understand this relationship and its implications on training design and outcomes. Future research also needs to examine how statistical package skills can be properly assessed incorporating this very likely dependency. Regardless, this study was the first to provide evidence of a relationship between statistical package skills and knowledge of statistics. This relationship will be important to control for in future studies that compare the effectiveness of different training strategies.

The second aim of this study was to investigate important advantages and disadvantages to implementing either of the training strategies into an introductory statistics course. This study looked at students' self-reported perceptions of training difficulty, training satisfaction, training anxiety and statistical package self-efficacy. Some instructors might be concerned that the EMT strategy might be more difficult for students leading to increased anxiety and lower self-efficacy. This may then lead to lower overall student satisfaction towards training. However, the results of this study failed to find any statistically significant evidence to support this concern. There were no significant differences between students' mean self-reported ratings of these outcomes.

The overall null findings of this study were surprising, but a number of limitations to the study and training design must be considered before drawing conclusions. EMT was hypothesised to have the greatest effect on adaptive transfer, but with 4 out of the 8 adaptive transfer tasks being removed due to online technical difficulties for post-training self-assessment, the exact effect of EMT on adaptive transfer at post-training remains to be seen. It is difficult to determine what would have happened if the error did not occur, but it would be safe to assume that the inclusion of four more adaptive transfer tasks would have introduced more variability in adaptive transfer scores and made it easier to detect differences between strategies if those differences existed.

In terms of the study design, this experiment was un-blinded. While students were never explicitly made aware of the nature of this study, it is highly probable that students became aware of the difference between strategies as the semester progressed. The supervisor was also un-blinded to the nature of the strategies. While it is difficult to speculate the exact influence this might have had on the results, the potential for bias cannot be ruled out. However, this type of experimental control is always going to be difficult to achieve in real-world educational research.

The major strength of this study, ecological validity, i.e. embedding the evaluation of EMT into a real introductory statistics course, was also its greatest limitation. Due to limited laboratory availability, training was scheduled on a fortnightly basis for each group. This meant that students had only a minimum estimated training time of four hours with *SPSS* before taking the self-assessment tasks. Given the large time intervals between training and the relative shortness of training, it is possible that the effects of training were interrupted and poorly consolidated. Future studies need to provide more frequent and consistent training throughout a course.

The training laboratory sessions were compulsory, but a large number of students missed laboratory sessions on a regular basis. This raised issues with training compliance. Due to ethical reasons, these students were permitted to attend laboratory sessions of the opposite strategy or complete the laboratory sessions in their own time. However, these students still received their respective strategy's instructions as the laboratory sessions were delivered through an online learning system which based laboratory session instructions (GT vs. EMT) on their allocated strategy. The results of the statistical models predicting training transfer performance at post-training found that non-compliance with the self-assessment, i.e. doing the self-assessment outside of the designated laboratory session, was associated with higher self-assessment scores. Non-compliant students probably did not stick to the self-assessment time limit or received help from peers who had already completed the self-assessment tasks. As attendance was recorded at all laboratory sessions, controlling for measures of training adherence and self-assessment compliance in the statistical models have at least partially taken these limitations into account. However, future research could benefit by ensuring students remain blinded and are given extra incentive to attend allocated laboratory sessions.

The laboratory sessions were scheduled for one hour. While the training was designed to fit within this time period, anecdotally many students reported feeling under time pressure which resulted in them rushing through laboratory sessions and using guesswork to get the laboratory sessions done in the designated time. It is possible that time constraints negatively impacted the EMT strategy and violated the error framing instructions. Under time constraints, it would be very difficult for a student to view errors as anything else but a waste

of time. While the availability of computer laboratories was outside the control of the researchers, a possible solution to this problem would be to provide further training opportunities so that students had adequate time to work through training material.

All training was graded in terms of satisfactory completion and students were allowed multiple attempts at the training laboratory sessions and self-assessment tasks. This feature of training may have resulted in unmotivated students not expending their greatest effort on self-assessment tasks. Instead, they may have done just enough to attain a level of satisfactory completion. The issues of low incentive may have masked a participant's true ability on the self-assessment tasks. While randomisation provided some level of protection against this issue biasing a particular strategy, in the future, assessment that better engages students in demonstrating their ability to operate a statistical package should be used.

There were also a number of important limitations related to the delivery of training strategies and the assessment of statistical package training transfer. While the researchers of this study were familiar with active learning strategies, this was the first time EMT was implemented for statistical package training at the study's institution. It was also the first time, to the authors' knowledge, that statistical package adaptive training transfer outcomes were formally assessed and reported in the literature. As such, many aspects of this study required the adaptation of methods and measures used in previous research. Only one study by Dormann and Frese (1994) related specifically to statistical package training. However, due to the age of this study, the absence of a specific mention of adaptive transfer and implementation of a one off training session outside of a statistics course, the Dorman and Frese experiment provided only a limited insight into the delivery of EMT and assessment of training transfer outcomes. Therefore, the delivery and assessment of training transfer required careful evaluation and reflection.

The results of the manipulation checks brought the validity of the EMT training strategy into question. If this study implemented EMT successfully then, when compared to participants in GT, participants in the EMT strategy would be hypothesised to self-report more metacognitive activity, evidence of exploratory behaviour, positive attitudes towards making errors and better emotional control. The only difference observed between strategies on the manipulation checks was for the use of step-by-step instructions. While the EMT group scored significantly lower, they still had a highly positive average level of agreement. This rating seemed too high assuming minimal instruction had been used correctly in the EMT strategy. It's likely that participants in the EMT strategy perceived the sequential delivery of exercises during training and the provision of training hints as providing guidance similar to step-by-step instructions. Regardless, it is clear from the results of these manipulation checks that there was a problem with the validity of the EMT strategy.

The self-assessment tasks used as measures of training transfer outcomes were also limited. As there was no literature to base the design of these tasks on, their validity as measures of analogical and adaptive transfer for statistical package training only extends to face validity. The strong relationship between statistical knowledge and training transfer suggests that less dependent methods need to be explored in order to get a more valid measure of a student's ability to operate a statistical package. The degree to which the self-assessment tasks captured analogical versus adaptive transfer was also an issue. Adaptive transfer is likely to be demonstrated by what students do spontaneously when working on their own statistical analysis problems outside of training. The degree to which this ability was captured using the self-assessment tasks used in this study was questionable. Future research on the assessment

of statistical package training transfer is needed so that these outcomes can be reliably and validly measured in the future.

After a critical analysis of the results, manipulation checks and methods, it is clear that further research is needed before a clear conclusion is reached about the relative merit of EMT over GT for statistical package training. While this study may have been unsuccessful in detecting the true effect of EMT, it does provide a valuable foundation to support future studies in this fertile area of statistics education. Specifically, future studies need to address the validity of implementing EMT for statistical package training in introductory statistics courses and assessing training transfer using reliable and valid measures. Future research in this area needs to continue to address ecological validity. The literature is already flooded with studies demonstrating the external validity of the superiority of EMT over GT in highly controlled studies (Keith & Frese, 2008). However, until the superiority of EMT can be demonstrated in real-world introductory statistics courses, EMT cannot be recommended over GT. It remains to be seen whether "less guidance is more" when it comes to training students how to use statistical packages in introductory statistics courses.

References

Baglin, J., Da Costa, C., Ovens, M., & Bablas, V. (2011). An experimental study comparing two different training strategies on how to use statistical software packages in an introductory statistics course. In M. Sharma, A. Yeung, & T. Jenkins et al. (Eds.), *Proceedings of the Australian Conference on Science & Mathematics Education (17th Annual UniServe Science Conference): Teaching for diversity – Challenges & strategies*, 28 – 30 September 2011 (pp. 162-168). Melbourne, Victoria. Retrieved December 25, 2011, from http://ojs-prod.library.usyd.edu.au/index.php/IISME/article/view/4805

Baglin, J., & Da Costa, C. (2012) Students' thoughts and perceptions of training to use statistical packages in introductory statistics courses: A qualitative study. In *Proceedings of the 8th Australian Conference on Teaching Statistics (OZCOTS): Statistics Education for Greater Statistics*, 12 – 13 July 2012. Adelaide, South Australia. Retrieved July 26, 2012, from http://opax.swin.edu.au/~3420701/OZCOTS2012/OZCOTS2012_BaglinJ_Final_paper.pdf

Bell, B. S., & 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.

Ben-Zvi, D. (2000). Toward understanding the role of technological tools in statistical learning. *Mathematical Thinking and Learning*, *2*, 127-155. Routledge.

Ben-Zvi, D., & Garfield, J. (2005). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15). New York: Kluwer Academic Publishers.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell & M. E. M. Mussen (Eds.), *Handbook of child psychology: Cognitive development* (Vol. 3, pp. 77-166). Harvard University Press.

Chance, B. L., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). The role of technology in improving student learning of statistics. *Technology Innovations in Statistics Education, 1*, 1 - 26. Retrieved September 12, 2012, from http://escholarship.org/uc/item/8sd2t4rr

Dormann, T., & Frese, M. (1994). Error training: Replication and the function of exploratory behaviour. *International Journal of Human-Computer Interaction, 6*, 365-372.

Finney, S., & 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., & 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., & Thiemann, P. (1991). Errors in training computer skills: On the positive function of errors. *Human-Computer Interaction, 6*, 77-93.

GAISE. (2005). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) college report*. American Statistical Association. Retrieved August 25, 2012, from http://www.amstat.org/education/gaise/GAISECollege.htm

Garfield, J., & Ben-Zvi, D. (2005). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 397-409). New York: Kluwer Academic Publishers.

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education, 10*. Retrieved August 25, 2012, from http://www.amstat.org/publications/jse/v10n2/garfield.html

Hesketh, B. (1997). Dilemmas in training for transfer and retention. *Applied Psychology, 46*, 317-339.

Ivancic, K., & Hesketh, B. (1996). Making the best of errors during training. *Training Research Journal, 1*, 103-125.

Kanfer, R., Ackerman, P. L., & Heggestad, E. D. (1996). Motivational skills & self-regulation for learning: A trait perspective. *Learning and Individual Differences, 8*, 185-209.

Keith, N., & 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., & Frese, M. (2008). Effectiveness of error management training: A meta-analysis. *Journal of Applied Psychology, 93*, 59-69.

Keith, N., Richter, T., & 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., & Tammen, 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.

Rybowiak, V., Garst, H., Frese, M., & Batinic, B. (1999). Error Orientation Questionnaire (EOQ): Reliability, validity, and different language equivalence. *Journal of Organizational Behaviour, 20*, 527-547.

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

Wood, R., Kakebeeke, B., Debowski, S., & Frese, M. (2000). The impact of enactive exploration on intrinsic motivation, strategy, and performance in electronic search. *Applied Psychology, 49*, 263-283.