Comparison of Intellectus Statistics and Statistical Package for the Social Sciences

Differences in User Performance Based on Presentation of Statistical Data

Allen C. Chen(✉), Sabrina Moran, Yuting Sun, and Kim-Phuong L. Vu

California State University, Long Beach, Long Beach, CA 90840, USA
achen31@gmail.com, sabrina.n.moran@gmail.com,
debby130403@gmail.com, kim.vu@csulb.edu

Abstract. Data-to-text systems create reports using natural language to simplify the presentation of complex data. Intellectus Statistics (IS) is a cloud-based statistical analysis software that provides users with output displayed in American Psychological Association (APA) narrative format. Statistical Package for the Social Sciences (SPSS) is another statistical analysis software; however, SPSS output is mainly presented numerically with tables and graphs. The purpose of this study was to compare the effectiveness and efficiency of using IS and SPSS to conduct and interpret analyses. An output presented in narrative format could be beneficial to students learning statistics who may have difficulty interpreting results. Overall, accuracy scores and time on task for the two software were not significantly different. Perceived usability and ease of use ratings for IS were significantly higher compared to SPSS. On the other hand, ratings of perceived usefulness were not significantly different between the two software. Results also suggested that participants preferred IS and felt more confident in conducting statistical analyses when using the software. Though there was no significant difference in task accuracy between the two software, data-to-text output helped students with interpreting assumptions for analyses and formatting written results.

Keywords: Data-to-text systems · Data interpretation
Visual display of information

1 Background

Statistics is the science of collecting, analyzing, and interpreting data. To produce accurate reports and interpretations of data, appropriate steps must be taken prior to analyses. Some common pitfalls that lead to inaccurate results include incorrect analysis choice, violations of assumptions, and incorrect interpretation of results. As the demand for statistical knowledge has grown in the age of big data, employment for statisticians is projected to increase 34% between 2014 and 2024 [9]. Furthermore, a report by Wasserstein [10] noted that completion rate for bachelor's degrees in statistics outpaced all other STEM disciplines in the four years prior to the paper's publication.

D. Harris (Ed.): EPCE 2018, LNAI 10906, pp. 389–402, 2018.
https://doi.org/10.1007/978-3-319-91122-9_32

1.1 Intellectus Statistics

Intellectus Statistics (IS) is a statistical analysis software created by Statistics Solutions. IS provides a cloud-based platform where users are presented with output displayed in APA narrative format after conducting statistical analyses. The data-to-text output also reports on the assumptions of analyses and indicates any violations. Past research has shown that students in introductory statistics courses have difficulty writing conclusions based on results from their data analysis [5]. Data-to-text systems simplify the presentation of data by using natural language to generate reports [3], which may be beneficial to users with limited statistical background.

Potential users of IS include anyone seeking to conduct statistical analyses. With an APA formatted output, the target user group is assumed to be undergraduate and graduate students in the social sciences. The results, which includes graphs and tables, are presented in a format commonly used in those disciplines. The interface was designed to be simple and easy to use, and an output in the form of a narrative would appeal to users with limited statistical background. In addition to conducting different types of analyses, the software also assists with data cleaning and data visualizations.

The ability to access the software and saved data from any computer with an internet connection makes IS attractive to users who rely on public computers or have limited storage space. While being cloud-based has its advantages, the platform is constrained by maintenance of working code, product support, and the version update process. Any potential issues with the cloud-based software would affect all users.

1.2 Statistical Package for the Social Sciences

Statistical Package for the Social Sciences (SPSS) is another statistical analysis software. The original target users of SPSS were students and professors in the social sciences, but its use has grown to other fields such as the medical sciences. SPSS offers greater functionality compared to IS; however, the interface is less intuitive and often uses terms unfamiliar to novice users. Much of the output, presented numerically with tables, requires some statistical background for accurate interpretation.

Potential users of SPSS consist of individuals who need to conduct statistical analyses. SPSS is commonly used by and taught to psychology students at many educational institutions. Through the use of commands in the Syntax Editor, SPSS offers intermediate and expert users increased flexibility and efficiency of use. Another benefit, due to its wide use in academia, is the wealth of help documentation available on the web. SPSS is often daunting for novice users due to the amount of information, options, and unfamiliar terminology. The program can be downloaded and installed on individual computers. Software updates are not automatic and are based on user preference. A cloud-based version of SPSS is also available, but its use could result in the same problems discussed earlier with cloud-based programs.

1.3 Purpose of Current Study

The purpose of this usability evaluation was to compare IS and SPSS' effectiveness and efficiency for students who have some prior experience conducting statistical analyses. An output presented in narrative format could be beneficial to students who have trouble interpreting results by providing more contextual information alongside the numbers. The students in the current study completed tasks involving data entry, data cleaning, conducting statistical analyses, interpretation of assumptions, and interpretation of results. Task accuracy, time on task, and perceived usability was measured and compared. Perceived usability was measured using the System Usability Scale (SUS). Perceived ease of use and usefulness was measured using the extended Technology Acceptance Model (TAM 2).

2 Method

2.1 Metrics

Time on task, task accuracy, and task completion rate were used to assess efficiency and effectiveness. Time on task was calculated by measuring the length of time that participants spent on each subtask. Length of time began when the participant started moving the computer mouse for navigation and ended when the participant verbally confirmed completion of the subtask. Combining the times for each subtask resulted in the completion time for the overall task. Shorter task times would indicate increased efficiency [1].

Task accuracy was determined using a scoring rubric. Points were awarded for correctly running analyses and correct interpretation of output. Scoring for interpretation was also dependent on including necessary information for APA formatted results. Points for originality were awarded based on Turnitin Similarity scores, which indicated a percent of matching content with other sources. The same scoring rubric was used for both software and tasks. Task accuracy was scored by two independent raters and high inter-rater reliability was found (Cronbach's $\alpha = 0.87$ to 0.96). Higher task accuracy would indicate increased task success and increased effectiveness [1].

Overall perceived usability was obtained using the System Usability Scale (SUS). The SUS is a self-reported 10-item Likert scale questionnaire that measures overall perceived reliability. According to previous research, the SUS is a highly robust tool for measuring perceived usability [2]. Possible perceived usability scores range from zero to 100. Scores under 50 indicate unacceptable usability, while scores above 70 indicate acceptable usability [1]. Scores ranging from 51 to 70 indicate marginal acceptability.

Perceived usefulness and ease of use was obtained using scales from the extended Technology Acceptance Model (TAM 2). Perceived usefulness and ease of use were reduced to four Likert-scale items ranging from one to seven in the extended model. Across studies and time periods, the internal reliability for both perceived usefulness (Cronbach's $\alpha = 0.87$ to 0.98) and ease of use (Cronbach's $\alpha = 0.86$ to 0.98) in the TAM 2 were high [8]. Previous research has shown that perceived usefulness is a strong determinant of intentions to use, while ease of use is not a significant predictor [4]. Higher scores represent higher perceived usefulness and ease of use.

2.2 Materials and Equipment

The usability evaluations took place at the Center for Usability in Design and Accessibility. Participants completed user testing in a room separated from the researcher using a one-way window. Participants first completed an informed consent form then completed the tasks on a Dell desktop computer running on the Windows 10 operating system. Tasks that required the use of SPSS were completed using SPSS Version 23. IS was accessed in a Google Chrome browser. Microsoft Word was used by participants to type the answers for tasks. The participants' screen was recorded using Morae, and Google Hangouts was used for communication between the researcher and participants during the study. After each task, the participants were given surveys to measure perceived usability, ease of use, and usefulness. A post-test questionnaire was also given to participants regarding their preference of software.

2.3 Usability Testing

Tasks. Two tasks were developed. Each task consisted of a scenario to provide participants context for the current study. In the scenario, participants were told that they were students in a statistics course who needed to complete homework. The tasks were designed to simulate problems that students would typically encounter in introductory and intermediate statistics courses in the behavioral sciences.

The study used a within-subjects design and measured the performance of both software products for each participant. In addition, the software products and order of tasks were counterbalanced to account for order effects like learning and fatigue. Participants also tended to be somewhat experienced using SPSS, but had no previous experience with IS. Directions for both software were provided with the tasks to reduce performance bias associated with different levels of experience with the two software products.

The two tasks each contained different datasets. The first task involved a researcher interested in Body Mass Index who wanted to evaluate the effectiveness of an exercise program. The second task involved a researcher interested in standardized test scores and how they relate to overall school GPA. Each task was then broken down into six subtasks. The subtasks involved (1) data entry, (2) data cleaning, (3) data visualization, (4) descriptive analysis, (5) independent t-test, and (6) linear regression. Answers to the subtasks were typed in a Microsoft Word document.

Participants. Participants were 12 self-selected students ($N = 12$) who responded to recruitment flyers posted around the psychology building at California State University, Long Beach, and were compensated 15 dollars ($15 USD). There were five male and seven female participants. Five participants were pursuing an undergraduate degree in psychology. Seven participants were graduate students in the psychology department. The participants' mean overall GPA in statistics courses was 3.47 ($SD = 0.39$) out of a four-point scale. Undergraduate students had the same mean GPA ($M = 3.47$, $SD = 0.37$) as graduate students ($SD = 0.46$). To complete the tasks for both software products, participants needed previous experience with data analysis and writing APA formatted results. A prerequisite for participants in this study was the completion of intermediate statistics at

California State University, Long Beach. Information about participants' perceived level of understanding of statistics was obtained using a multiple-choice question with the possible answers of beginner, intermediate, or expert. Most users ($n = 9$) reported having an intermediate level of understanding of statistics.

Participants' experience and comfort with statistical analyses and writing results were measured using scales that ranged from one to five. From very inexperienced (1) to very experienced (5), participants indicated that they were somewhat experienced with performing data analyses ($M = 3.50$, $SD = 0.80$). From very uncomfortable (1) to very comfortable (5), participants indicated that they were between neutral and somewhat comfortable with writing statistical results ($M = 3.42$, $SD = 1.08$). Experience with statistical analyses and comfort with writing statistical results were strongly correlated, $r(10) = .89$, $p < .001$. Regarding experience with statistical software, participants had no previous experience using IS. Participants reported being somewhat experienced using SPSS ($M = 3.67$, $SD = 0.78$) and somewhat comfortable using SPSS ($M = 3.50$, $SD = 1.00$).

Procedure. Participants were first given the general background of the study and statistical software. Participants were then given time to explore the software prior to starting the task. For IS, participants were shown a brief video overview then given two minutes to explore the software. For SPSS, participants were given two minutes to explore the software prior to the task. Participants were not shown a video overview for SPSS as all participants had previous experience with the software.

After exploration of the software, participants began completing the tasks. Answers to the subtasks were entered in a Microsoft Word document. After finishing the first task, participants completed the SUS along with the surveys for perceived usefulness and ease of use. The same procedure applied for the second task. Participants were compensated after the questionnaires for the second task had been completed. The entire study lasted approximately 90 min.

3 Results

3.1 Analyses

Statistical analyses for the usability evaluation were conducted using SPSS Version 25. Task accuracy and time on task were analyzed using mixed design analyses of variance (ANOVA). Assumptions were inspected and violations were not found. The assumption of sphericity was not evaluated as there were only two levels for within-subjects variables. The results showed that completion rate for the regression subtask was lower than 70% for both IS and SPSS. Therefore, data for the regression subtask was not included in the analyses. Performance measures for effectiveness include task completion rate and task accuracy. Time on task was used as the performance measure for efficiency.

3.2 Task Completion Rate

All participants ($N = 12$) finished the first five subtasks for both software. For the linear regression subtask, eight participants completed the problem using SPSS ($n = 8$) and six participants completed the problem using IS ($n = 6$) (Fig. 1). Most participants did not finish the tasks due to a time constraint with the length of test session. The task completion rate was 67% and 50% for SPSS and IS, respectively. Consequently, data for the linear regression subtask was not used in the analysis on task accuracy and time on task.

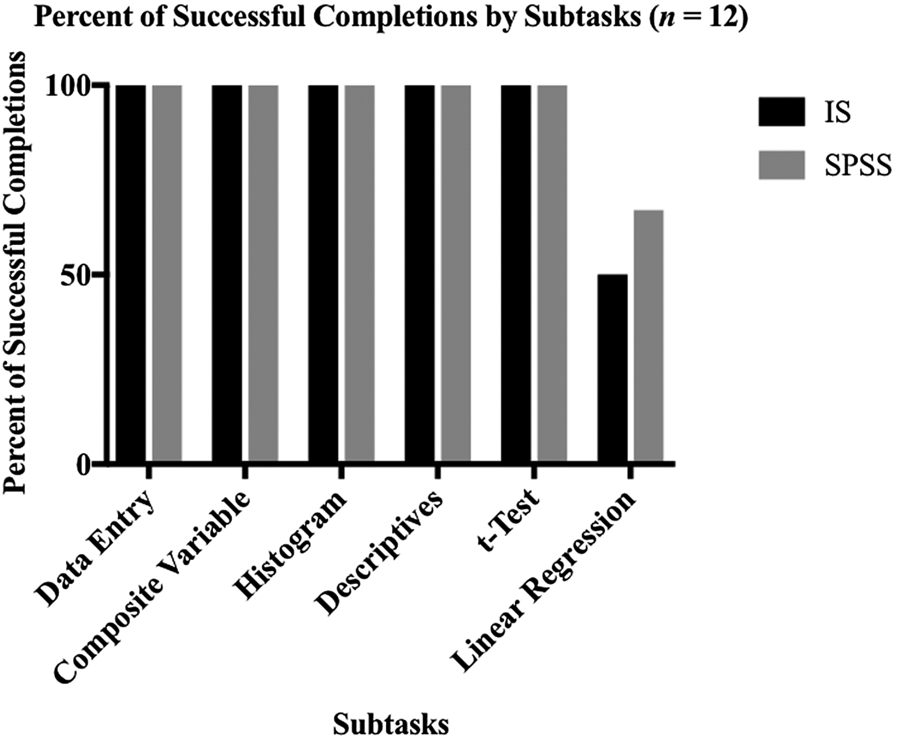

Fig. 1. Percent of successful completions for different subtasks and software.

3.3 Task Accuracy

Results indicate there was no significant difference in accuracy scores based on type of software used, $F(1, 8) = 0.38, p = .557$ (Fig. 2). The mean overall accuracy scores were 77.17 ($SD = 6.53$) and 78.83 ($SD = 7.91$) for SPSS and IS, respectively. There was also no significant difference in accuracy scores between the two tasks, $F(1,8) = 0.02, p = .901$, or order of software used, $F(1, 8) = 0.22, p = .651$. The mean accuracy scores were 78.16 ($SD = 10.35$) and 75.91 ($SD = 6.62$) for Task 1 and Task 2, respectively.

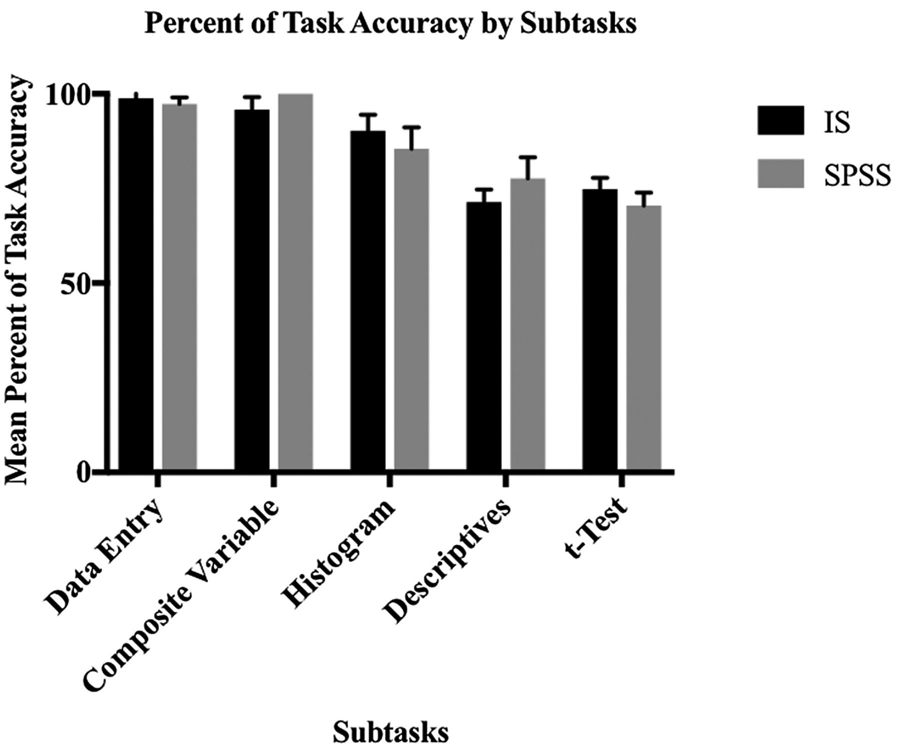

Fig. 2. Mean percent of task accuracy for different subtasks and software.

3.4 Originality Scores

Originality scores for answers using IS were not significantly different to originality scores for answers using SPSS, $t(11) = 1.99$, $p = .072$ (Fig. 3). However, the average percent of matching content to other sources was higher for IS ($M = 32.08$, $SD = 36.23$) compared to SPSS ($M = 8.17$, $SD = 22.06$). There were three answers for IS and one answer for SPSS that exceeded 70% matching content.

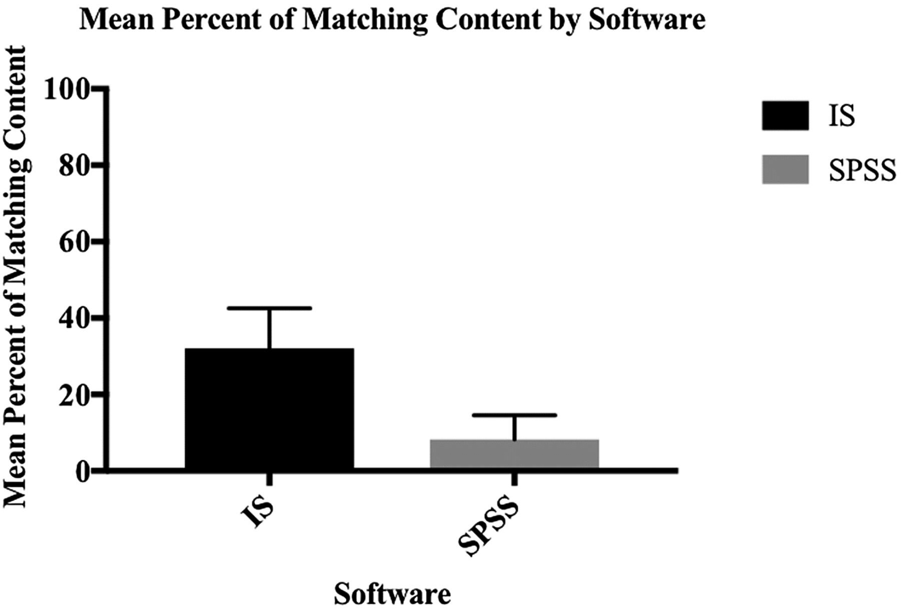

Fig. 3. Mean percent of content in participants' answers that matched with other sources.

3.5 Time on Task

Results for time on task (in seconds) indicate there were no significant differences between IS (*M* = 1697.58, *SD* = 1592.08) and SPSS (*M* = 1592.08, *SD* = 239.14), *F*(1, 8) = 1.59, *p* = .243 (Fig. 4). There was also no significant difference in time on task between Task 1 and Task 2, *F*(1, 8) = 0.15, *p* = .707. However, a significant difference was found in mean completion time for order of software used, *F*(1,8) = 8.16, *p* = .021, $\eta_p^2 = 0.51$. Participants who used IS first spent a significantly longer time on tasks (*M* = 1823.58, *SD* = 355.34) compared to participants who used SPSS first (*M* = 1466.08, *SD* = 192.10).

There was also an interaction for time on task between type of software used and order of software used, *F*(1,8) = 10.04, *p* = .013, $\eta_p^2 = 0.56$. Participants who used IS for their first task spent a shorter time on the second task (*M* = 1638.17, *SD* = 251.12) compared to the first task (*M* = 2009.00, *SD* = 363.62). Similarly, participants using SPSS for their first task also spent a shorter amount of time on their second task (*M* = 1386.17, *SD* = 90.56) compared to the first task (*M* = 1546.00, *SD* = 240.11). However, the difference in mean time on task between the two software was significantly larger for participants who used IS first compared to those who used SPSS first.

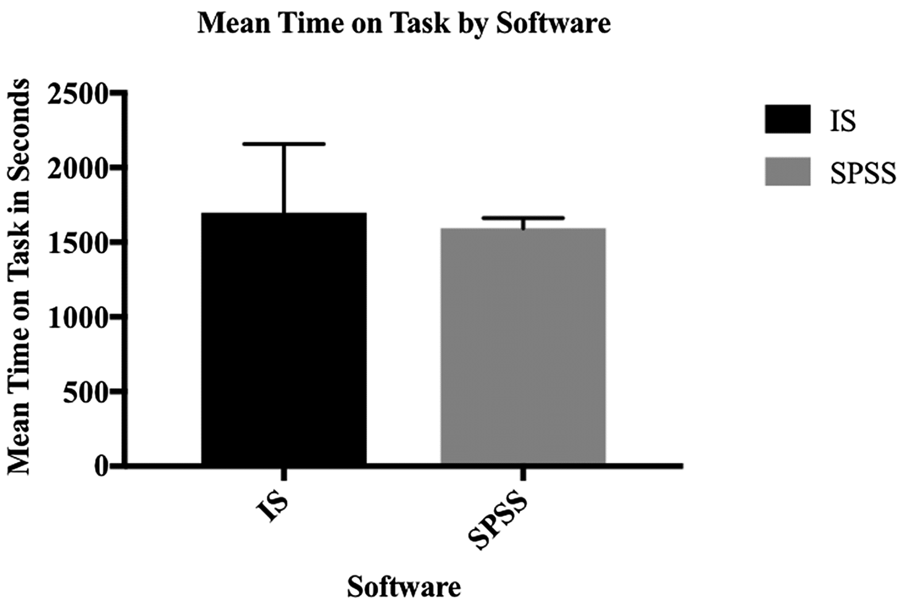

Fig. 4. Mean time on task in seconds for IS and SPSS.

3.6 Perceived Usability – SUS

SUS scores indicated that participants ($N = 12$) felt IS was significantly more usable compared to SPSS, $t(11) = 5.32$, $p < .001$. SUS scores for IS ($M = 83.33$, $SD = 11.74$) is considered to be at an acceptable level of usability [1]. SUS scores for SPSS ($M = 47.50$, $SD = 15.67$) were significantly lower and is considered to be at an unacceptable, but close to a marginal, level of usability (Fig. 5).

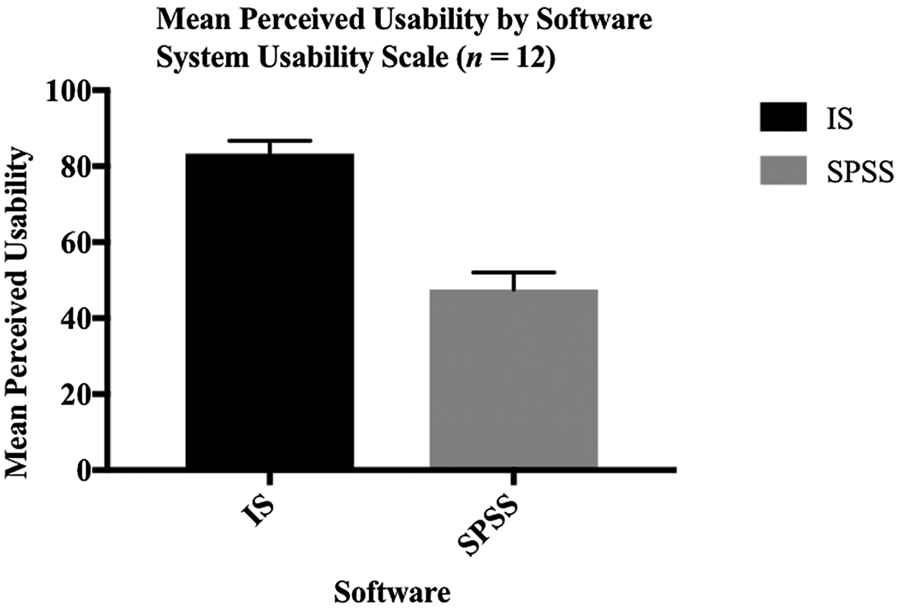

Fig. 5. Mean perceived usability scores for IS and SPSS.

3.7 Perceived Ease of Use and Usefulness – TAM 2

Results showed that participants felt that IS ($M = 5.88$, $SD = 0.84$) was significantly easier to use compared to SPSS ($M = 3.71$, $SD = 0.88$), $t(11) = 5.38$, $p < .001$ (Fig. 6). However, it was indicated that participants did not feel that IS ($M = 5.60$, $SD = 0.96$) was more useful than SPSS ($M = 5.06$, $SD = 1.02$), $t(11) = 1.34$, $p = .206$ (Fig. 7). Perceived ease of use for SPSS was not correlated with perceived usefulness, $r(10) = .16$, $p = .61$. The association between perceived ease of use and perceived usefulness for IS was also not significant, $r(10) = .57$, $p = .055$.

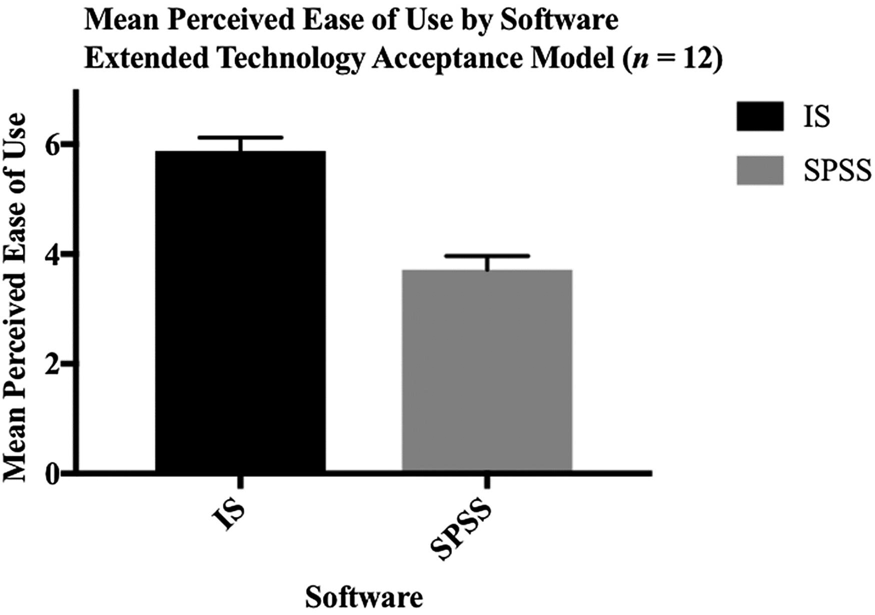

Fig. 6. Mean perceived ease of use rating for IS and SPSS.

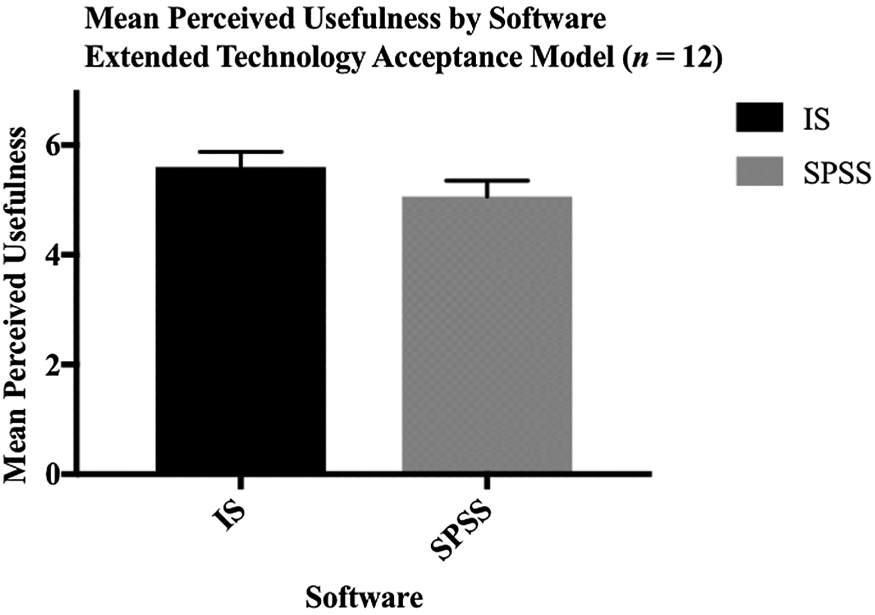

Fig. 7. Mean perceived usefulness rating for IS and SPSS.

3.8 Fear and Confidence Towards Conducting Statistical Analyses

Participants completed a post-test questionnaire after completing both tasks. Preference of statistical analysis software was measured using a scale with one indicating IS and seven indicating SPSS. Participants tended to prefer IS over SPSS (*M* = 2.92, *SD* = 1.31) (Fig. 8). Participants were also asked to rate if they felt that the statistical software reduced their overall fear of statistics using a Likert scale. A score of one indicated strongly disagree and a score of seven indicated strongly agree. Participants felt near neutral (*M* = 4.75, *SD* = 1.28) for IS, while disagreeing with the statement for SPSS (*M* = 2.83, *SD* = 1.11). Using the same scale, participants were asked to rate their agreement with a statement asking if using the statistical software increased their confidence in conducting statistical analyses. Participants agreed with the statement for IS (*M* = 5.83, *SD* = 0.72), but felt near neutral for SPSS (*M* = 3.33, *SD* = 1.44).

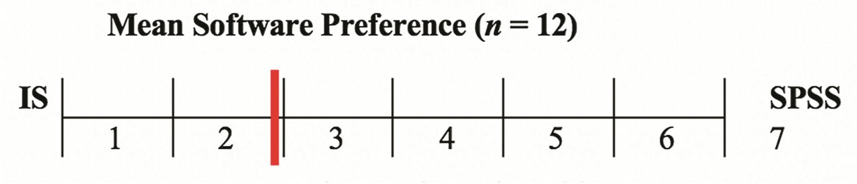

Fig. 8. Mean software preference of participants for IS and SPSS.

4 Discussion

The hypothesis of the current study was that IS would be more effective, efficient, and have higher usability compared to SPSS. To examine effectiveness, task accuracy was analyzed and there were no significant differences found. Overall, accuracy scores for IS and SPSS were fairly similar. One of the factors that led to higher effectiveness for SPSS when interpreting descriptive statistics is the amount of information provided in the output. While IS only provided the minimum and maximum values in the output, SPSS also provided the range. Many participants, despite having access to a calculator, did not calculate the difference between maximum and minimum values to obtain the range when using IS.

Another factor that influenced effectiveness of the software was the output presentation. The difference in accuracy scores for the t-test was largely due to poor performance in interpreting assumptions and including necessary information in the written results. The narrative format of IS output allowed users to easily obtain assumptions and check the format of their written results. SPSS output forced users to recall previous knowledge on assumptions and formatting requirements of APA results.

Scores for originality supported the conclusion that participants relied on the narrative format of IS output as a template when writing results. Though not statistically significant, IS had a higher mean percent of matching content compared to SPSS. Three participants using IS had written results with over 70% of their content matching with other online sources. Several participants also made a statement about wanting to copy and paste the IS output to their written answers. This indicates potential issues, such as plagiarism, for students using IS for coursework. Past research also suggests that elaborative processing, though requiring more attention and time, results in improved learning and memory compared to shallow processing [6]. Despite having an interface that is more difficult to use, the need to interpret SPSS output in a contextual manner may lead to more effective student learning compared to IS.

To examine efficiency, time on task was inspected and a main effect for the order of software used was found. There was also an interaction between type and order of software used. A potential reason for this effect was that participants, who were inexperienced with IS, took a longer time to read and locate relevant information in the narrative output in the first task. However, if the narrative output was presented second, participants would likely skim the output due to fatigue or time constraints with the test session.

Perceived usability and ease of use scores for IS were significantly higher compared to SPSS. IS has an interface that is easier for users with a limited background in statistics to understand. The interface is more simplistic than SPSS and less overwhelming to use. The left menu only contains four major tabs and the menu options uses terminology familiar to users. This finding was expected as IS was designed for and marketed to users with limited statistical knowledge.

On the other hand, perceived usefulness was not significantly different for IS and SPSS. It is important to note that perceived usefulness is a stronger predictor of actual system usage than perceived ease of use [4]. While IS is simpler and easier to use, SPSS has more functionality. Many of the participants were seniors and graduate students that were taking or had taken advanced statistics courses. Some features in SPSS commonly used by intermediate or expert level users, including syntax and recently recalled dialogs, are not available in IS and may have influenced perceived usefulness ratings.

Overall, the post-test questionnaire suggested that participants tended to prefer IS over SPSS. Participants also felt that IS increased their confidence in conducting statistical analyses. This is most likely due to the easy-to-use interface combined with a narrative output. Past research has investigated the influence of self-efficacy, anxiety, and self-confidence on mathematics achievement in school [7]. The results suggested that confidence is strongest non-cognitive predictor of academic achievement. Thus, it is possible that IS may be an effective supplemental teaching tool for students learning statistics. The research hypothesis that IS would be significantly more effective, efficient, and have higher usability was only partially supported. Results did not support a significant difference in effectiveness and efficiency between the two software; however, IS was found to have significantly higher perceived usability and ease of use.

4.1 Limitations

The tasks and scoring rubric used in the present study were constructed by researchers with instructional experience in statistics. However, the tasks and rubric should be constructed in conjunction with a statistics subject matter expert to improve ecological validity. The 90-min time limit of the study was problematic as users were not able to complete all the tasks, which resulted in missing data for the regression subtask. Most participants were also showing signs of fatigue after the first task. Similar studies in the future should be designed with shorter tasks to better account for participant fatigue. Participants for future studies should also have experience or training using IS. Ensuring that participants have comparable experience using both software would allow for the most accurate comparisons.

References

1. Albert, W., Tullis, T.: Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Newnes, London (2013)
2. Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. Int. J. Hum.-Comput. Inter. **24**(6), 574–594 (2008)
3. Gkatzia, D.: Content selection in data-to-text systems: a survey. arXiv preprint arXiv:1610.08375 (2016)
4. Guritno, S., Siringoringo, H.: Perceived usefulness, ease of use, and attitude towards online shopping usefulness towards online airlines ticket purchase. Procedia-Soc. Behav. Sci. **81**, 212–216 (2013)
5. McGrath, A.L.: Content, affective, and behavioral challenges to learning: students' experiences learning statistics. Int. J. Scholarsh. Teach. Learn. **8**(2), 6 (2014)
6. Schmeck, R.R.: Improving learning by improving thinking. Educ. Leadersh. **38**(5), 384–385 (1981)
7. Stankov, L., Morony, S., Lee, Y.P.: Confidence: the best non-cognitive predictor of academic achievement? Educ. Psychol. **34**(1), 9–28 (2014)
8. Venkatesh, V., Davis, F.D.: A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag. Sci. **46**(2), 186–204 (2000)
9. Understanding the 2014–24 projections: Career Outlook. https://www.bls.gov/careeroutlook/2015/article/projections-methodology.htm. Accessed 27 Nov 2017
10. Wasserstein, R.: Communicating the power and impact of our profession: a heads up for the next Executive Directors of the ASA. Am. Stat. **69**(2), 96–99 (2015)