Published in HFES 2006 proceedings

DRIVING SIMULATOR EXPERIMENTS: POWER FOR REPEATED MEASURES vs. COMPLETELY RANDOMIZED DESIGN

Birsen Donmez, Linda Boyle, John D. Lee
Department of Mechanical and Industrial Engineering
University of Iowa
Iowa City, Iowa

Driving simulator studies are usually designed to collect repeated measures on each participant. This design has different implications for the power of within and between-subject effects that needs to be recognized by researchers. The power for between-subject variables decreases when additional measures are collected on the within-subject variables. However, the power for the main and interaction effects of the within-subjects variables increases as more observations are collected on one participant. If the main interest of the experiment is a between-subject effect, such as age, then a completely randomized design can provide the same power with fewer participants. Through a case study, this paper demonstrates how the power changes between a repeated measures design and a completely randomized design.

INTRODUCTION

Driving simulators are widely used in human factors and transportation research to explore issues in driving behavior, such as the assessment of driving performance given medical impairments (Findley et al., 1989; Rizzo, McGehee, Dawson, & Anderson, 2001), age effects (Brouwer, Waterink, Van Wolffelaar, & Rothergatter, 1991; Syzlek et al., 1995), and the design of new transportations systems ranging from roadway infrastructure to in-vehicle systems (Boyle & Mannering, 2004; Donmez, Boyle, & Lee, in press).

Like much behavioral research, an important issue with driving simulator studies is the limited number of power analyses that appear to be performed or reported (Cohen, 1990). Power analyses are important because they are needed to determine the sample size required to detect an effect before a study is conducted; to obtain the power of a study that was already conducted; and to assess the size of effect that could be reliably detected by a particular study. Power analysis can be performed *a priori* and post-hoc. *A priori* power analysis would lead to a better designed experiment by determining the number of subjects needed to obtain statistically significant results if there is an effect. *A priori* analysis requires estimated variances that would result from the experiment. Therefore, this analysis is approximate but very useful in determining the number of subjects required before a study begins. Generally, 80% power is targeted in determining the sample size. Power of 80% means that rejecting the null when there is an effect is four times more likely than not rejecting. Larger power is of course desirable, but the rate of increase in power for each additional participant is low when the power is large. Therefore, the benefit of the additional power over 80% usually does not match the cost of the additional participants. Post-hoc analysis can be performed to assess if the non-significance is due to lack of power and can provide insights on the experimental results. Further detail on performing power analysis can be found in Cohen (1988) and Murphy & Myors (2004). Surprisingly few researchers consider statistical power or effect sizes in reporting statistical results, a phenomenon that can undermine the ability of other researchers to properly interpret experimental results (Meehl, 1978; Vicente & Torenvliet, 2000).

Driving simulator experiments are generally designed as repeated measures studies where one subject goes through different experimental conditions (Horrey & Wickens, 2004; Rizzo et al., 2001; Tsimhoni & Green, 2001). Alternatively, in a completely randomized design, one subject would only be placed in one experimental condition. Compared to completely randomized design, repeated measures provide many advantages. For example, repeated measures design requires fewer participants and also increases the power of repeated variables by decreasing the error variance for these terms (Bradley & Russell, 1998; Murphy & Myors, 2004). Despite its many advantages, repeated measures experimental designs are more complex. When compared to the completely randomized design, repeated measures design requires more participants to provide the same power to test the between-subject effects.

It is important to design an experiment with the scientific question in mind. If the between-subject variable is of crucial interest to the experiment then a repeated measures design is not the best experimental design. Such a research question benefits from a completely randomized design. However, if the primary research question concerns the effects of the within-subject variable and how these effects change with different levels of the between-subject variable (i.e. interaction

term) then a repeated measures design provides similar power with fewer participants. Through a case study, this paper demonstrates how the experimental power differs for a repeated measures design compared to a completely randomized design. The statistical techniques applied in this case study are not novel by any means. However, as mentioned above few researchers consider power in reporting their results, not providing information to suggest that a power analysis was even conducted. Therefore, this case study aims to demonstrate how the design choice (i.e. repeated measures vs. completely randomized design) and the associated power affect experimental results and are very important to consider. This paper illustrates the power implications associated with common experimental designs employed in human factors research. Specifically the advantages and disadvantages associated with repeated measures design is demonstrated.

METHOD

The power of an F-test for analysis of variance increases with increasing sample size, significance level and treatment effects, and decreasing experimental error (Bradley & Russell, 1998; Cohen, 1988). However, the experimental design can also affect the power of F-tests performed on treatments. As mentioned in the introduction, repeated measures designs have both negative and positive impacts on experimental power. The measures collected on the between-subject variables are often correlated. As correlated measures are collected, the error variance for the within-subject variables and the interaction term decreases. This is due to the estimation and removal of systematic subject effects that are treated as sources of variation rather than as error. Also, when compared to a completely randomized design, there are fewer error degrees of freedom for the within-subject variable and the interaction term. Therefore, repeated measures design increases the power for the main and interaction effects for the within-subject variables. However, the inclusion of repeated measures increases the error variance of the between-subject terms, degrading the power for these terms.

If the treatment effects are assumed to be the same, the relationship between the effect size of a completely randomized design of two factors, A (having p levels) and B (having q levels), and the effect sizes of a repeated measures design with A as the between-subject and B as the within-subject variable are (equations are from Bradley & Russell (1998)):

$$f_{A(\mathrm{RP})} = \frac{f_{A(\mathrm{CR})}}{\sqrt{1 + (\mathrm{q} - 1)\rho}}$$

$$f_{B(\mathrm{RP})} = \frac{f_{B(\mathrm{CR})}}{\sqrt{1 - \rho}} \qquad (1)$$

$$f_{AB(\mathrm{RP})} = \frac{f_{AB(\mathrm{CR})}}{\sqrt{1 - \rho}}$$

where f: effect size, RP: Repeated measures design, CR: completely randomized design, and ρ: correlation between repeated measures.

Case Study

This paper presents a case study based on a driving simulator experiment that was conducted to assess the effects of driver distractions as well as the effectiveness of different distraction mitigation strategies. Details about the experiment can be found in Donmez et al. (in press). An overall summary of the methods and experimental design is provided to aid the reader with the definitions used in the results section. Data from sixteen middle-aged (range: 35 to 55; mean: 45, s.d.: 17.1) and twelve older (range: 65 to 75; mean: 69, s.d.: 11.3) drivers were collected. The experiment was a 2x7 repeated measures design with age as a between-subject variable (2 levels), and strategy as a within-subject variable (7 levels). The within-subject variable, strategy, included four different mitigation strategies and three baseline conditions.

During each condition, participants were asked to follow a lead vehicle which braked periodically. Because minimum Time-to-collision (TTC) is an important measure in braking response, this variable is chosen to demonstrate power analysis for repeated measures design in simulator studies. TTC is defined as the distance between the participant and lead vehicles divided by the relative velocity (i.e., the time that a collision would occur if the vehicle were to proceed at constant speed). Minimum TTC is the shortest time-to-collision during a braking event if the participant were to continue in the same path at the same velocity, thus, an increase in this variable would indicate a safety benefit. Therefore, minimum TTC has been proposed and used as a crash-avoidance metric in driving studies (Minderhoud & Bovy, 2001; Vogel, 2003) and is associated with the collision likelihood (Lee, McGehee, Brown, & Reyes, 2002).

The results section presents the findings of the post-hoc power analysis conducted on minimum TTC for this repeated measures design. Using the same data and assuming that the treatment effects were the same, the power of a completely randomized 2 (age) x 7 (strategy) design was also calculated. For a completely randomized design, different people experience each treatment combination. This increases the error variance for the within-subject and the interaction terms, decreasing power. However, it also decreases the between-subject variability and hence increases power for the between-subject term. The following section demonstrates this effect.

RESULTS

Post-hoc power analysis was performed on the minimum TTC. The correlation between repeated measures was moderate ($\rho = 0.39$). The power for the within-subject effect (strategy) and the interaction (age x strategy) was quite high (Table 1). As expected, these terms resulted in significant findings. However, for the main age effect, the power was very low. The lack of power may explain the non-significant result for this variable. For this experiment, the high power for

strategy and age x strategy and the low power for age was a tradeoff in power that was considered necessary based on the experimental objectives (Donmez et al., in press). The experiment focused on how different strategies would affect driving performance and how different age groups will respond to these strategies (i.e. strategy and age x strategy interaction). The main factor of age, which assessed how middle-aged drivers' driving differs from older drivers' driving, was not the main focus of the study.

Table 1. The overall statistical significance and power for minimum TTC for repeated measures design at $\alpha = 0.05$

	Minimum TTC (for repeated measures design)				
	F-value	*p*	*f*	*n′*	$1-\beta$
Age	$F(1,26)=0.40$	*	0.05	14	0.05
Strategy	$F(6,138)=6.35$	*<0.05*	0.50	21	0.99
Age x Strategy	$F(6,138)=2.43$	*<0.05*	0.30	21	0.77

* not significant; *f*: effect size defined by Cohen (1988)

Figure 1 shows how the power would change if the sample size in each age group was increased assuming that the effect sizes are constant. Increasing the sample size from 14 drivers in each age group to 22 helps achieve almost 100% power for the age x strategy interaction. However, the power for the age factor is still very low even if the sample size in each age group is increased to 40, requiring a total of 80 participants. This is a result of the very small effect size that is observed for the age factor. To obtain 80% power for age, with 14 subjects in each age group, an effect size of 0.55 is needed.

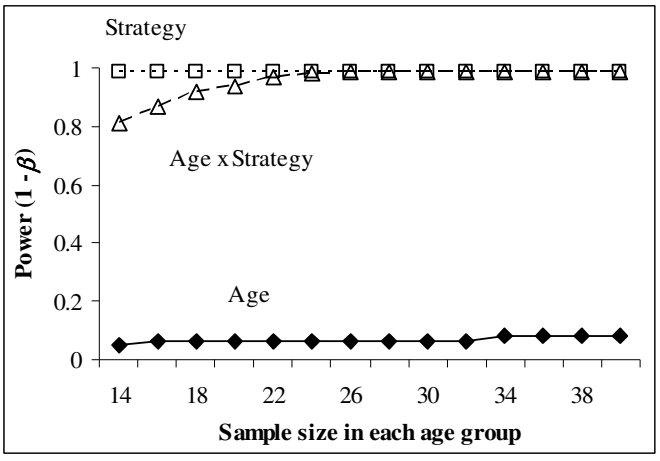

Figure 1. Power in repeated measures design as a function of sample size in each age group

If this experiment was a completely randomized design with the same levels of age and strategy, then the power for age would have been larger than 0.05. Table 2 shows the power values for an identical completely randomized design, where the treatment effects are assumed to be in the same magnitude. Even when the completely randomized design uses 192 participants (i.e. 7 x 28; 28 participants in each treatment group), the power of age x strategy is substantially lower than the power for the repeated measures design. The effects were assumed to be the same with the effects assessed in the repeated measures design for which the effect of age was extremely low. Therefore, the additional participants did not provide much of an increase in power for age in absolute terms (i.e. 0.03 increase in power), and thus the power for age is still low. However, in relative terms, there was a 60% relative increase from the power of repeated measures design (i.e. 0.05 to 0.08). This increase would have been more dramatic if the effect was larger.

Table 2. The hypothetical power for minimum TTC if the experiment was a completely randomized design at $\alpha = 0.05$ (calculated using equation 1)

	Minimum TTC (for completely randomized design)		
	f	*n′*	$1-\beta$
Age	0.10	14	0.08
Strategy	0.39	21	0.95
Age x Strategy	0.23	21	0.51

f: effect size defined by Cohen (1988)

Figure 2 shows how power would change in this experiment as a function of correlation between repeated measures. The graph shows the power of main and interaction effects as the correlation between measures increases. The power value for zero correlation ($\rho = 0$) is the power for an identical but completely randomized design. A dramatic change for age x strategy term is observed as the correlation increases. The power of strategy is high for the completely randomized design; therefore there is not much room for an increase in power.

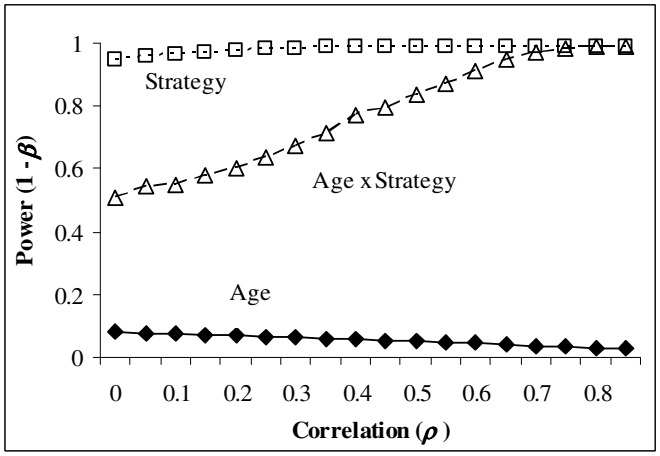

Figure 2. Power in repeated measures design as a function of correlation between measures (calculated using equation 1)

DISCUSSION

Many driving simulator studies are repeated measures design with one or more between-subject factors (e.g. age, gender). Repeated measures design is usually preferred because it is difficult to recruit and collect data from large number of participants. There is also great variability between drivers. Using a repeated measures design results in increased power for the within-subject variables and the corresponding interaction terms by decreasing their error variance. However, as the between-subject variable repeats the within-subject conditions, the error variance for the between-subject effect increases and thereby degrades the power for this main effect. There is a tradeoff in power which needs to be considered according to the objectives of the study. Without considering these objectives, one type of experimental design cannot be advocated over another.

A priori power analysis should be performed to determine the sample size to achieve the desired power. To perform such analysis, the variances for random effects (e.g. subject, random error) should be estimated. One way to obtain these estimates is to use data from similar past experiments. With varying experiments and simulator types this can be a challenge. Another way to estimate variances is to run a pilot study. A pilot study can provide better estimates but would require extra time and resources. Therefore, there is a tradeoff between accuracy and resources. Moreover, increasing sample size increases power. However, the rate of increase in power for the additional samples does not always justify the cost of running more participants. This presents another tradeoff between power and resources.

Using a case study, this paper demonstrates that a within-subject variable and its interactions have substantially higher power than a between-subject variable. Due to time and resource constraints, driving simulator studies usually do not involve a large number of participants (Breckenridge & Dodd, 1991; Brouwer et al., 1991; Reed & Green, 1999). It is likely that the between-subject terms in driving simulator studies that have repeated measures design will not generate significant results due to lack of power unless they have medium or large effect sizes. It is important to consider this artifact of repeated measures design in conducting driving simulator experiments. Neglecting this effect may lead researchers to mistakenly conclude that the between subject terms have little influence when in fact it is the opposite.

ACKNOWLEDGMENTS

This research was conducted as part of the SAVE-IT program under contract by Delphi Corporation and sponsored by the National Highway Traffic Safety Administration – Office of Vehicle Safety Research. The authors acknowledge the technical support and comments provided by Mike Perel of NHTSA and Mary Stearns of the Volpe Center.

REFERENCES

Boyle, L., & Mannering, F. (2004). Impact of traveler advisory systems on driving speed: some new evidence. *Transportation Research Part C, 12*, 57-72.

Bradley, D. R., & Russell, R. L. (1998). Some cautions regarding statistical power in split-plot designs. *Behavior Research Methods, Instruments, & Computers, 30*(3), 462-477.

Breckenridge, R. L., & Dodd, M. O. (1991). Locus of control and alcohol placebo effects on performance in a driving simulator. *Perceptual and motor skills, 72*(3 Pt 1), 751-756.

Brouwer, W. H., Waterink, W., Van Wolffelaar, P. C., & Rothergatter, J. A. (1991). Divided attention in experienced young and older driver: lane tracking and visual analysis in a dynamic driving simulator. *Human Factors, 33*, 573-582.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Second ed.). New Jersey: Lawrence Erlbaum Associates, Publishers.

Cohen, J. (1990). Things I have learned (so far). *American psychologist, 45*, 1304-1312.

Donmez, B., Boyle, L., & Lee, J. D. (in press). The impact of distraction mitigation strategies on driving performance. *Human Factors.*

Findley, L., Fabrizio, M., Knight, H., Norcross, B. B., LaForte, A. J., & Suratt, P. M. (1989). Driving simulator performance in patients with sleep apnea. *American review of respiratory diseases*(140), 529-530.

Horrey, W. J., & Wickens, C. D. (2004). Driving and side task performance: The effects of display clutter, separation, and modality. *Human Factors, 46*(4), 611-624.

Lee, J. D., McGehee, D., Brown, T. L., & Reyes, M. (2002). Collision warning timing, driver distraction, and driver response to imminent rear end collision in a high fidelity driving simulator. *Human Factors, 44*(2), 314-334.

Meehl, P. E. (1978). Theoretical risks and tabular asteriks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Minderhoud, M. M., & Bovy, P. H. L. (2001). Extended time-to-collision measures for road traffic safety assessment. *Accident Analysis & Prevention, 33*, 89-97.

Murphy, K. R., & Myors, B. (2004). *Statistical Power Analysis: A simple and general model for traditional*

and modern hypothesis testing. (Second ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Reed, M. P., & Green, P. A. (1999). Comparison of driving performance on-road and in a low-cost simulator using a concurrent telephone dialing task. *Ergonomics, 42*(8), 1015-1037.

Rizzo, M., McGehee, D., Dawson, J., & Anderson, S. (2001). Simulated car crashes at intersections in drivers with Alzheimer disease. *Alzheimer Disease and Associated Disorders, 15*, 10-20.

Syzlek, J. P., Pizzimenti, C. E., Fishman, G. A., Kelich, R., Wetzel, L. C., Kagan, S., et al. (1995). A comparison of driving in older subjects with and without age-related macular degeneration. *Archives of ophthalmology, 113*(8), 1033-1040.

Tsimhoni, O., & Green, P. (2001). *Visual demand of driving and the execution of display-intensive in-vehicle tasks.* Paper presented at the Human Factors and Ergonomics Society 45th Annual Meeting, Santa Monica, CA.

Vicente, K. J., & Torenvliet, G. L. (2000). The earth is spherical ($p<.05$): alternative methods of statistical inference. *Theoretical Issues in Ergonomics, 1*(3).

Vogel, K. (2003). A comparison of headway and time to collision as safety indicators. *Accident Analysis & Prevention, 35*, 427-433.