1-1-2004

Toward systematic control of cybersickness

Marshall B. Jones

Robert S. Kennedy

Kay M. Stanney
University of Central Florida

Marshall B. Jones
The Pennsylvania State University
College of Medicine, Hershey, PA
17033

Robert S. Kennedy
RSK Assessments, Inc., Orlando, FL
32803

Kay M. Stanney
University of Central Florida,
Orlando, FL 32816

Toward Systematic Control of Cybersickness

Abstract

Visually induced motion sickness, or "cybersickness," has been well documented in all kinds of vehicular simulators and in many virtual environments. It probably occurs in all virtual environments. Cybersickness has many known determinants, including (a short list) field-of-view, flicker, transport delays, duration of exposure, gender, and susceptibility to motion sickness. Since many of these determinants can be controlled, a major objective in designing virtual environments is to hold cybersickness below a specified level a specified proportion of the time. More than 20 years ago C. W. Simon presented a research strategy based on fractional factorial experiments that was capable in principle of realizing this objective. With one notable exception, however, this strategy was not adopted by the human factors community. The main reason was that implementing Simon's strategy was a major undertaking, very time-consuming, and very costly. In addition, many investigators were not satisfied that Simon had adequately addressed issues of statistical reliability. The present paper proposes a modified Simonian approach to the same objective (holding cybersickness below specified standards) with some loss in the range of application but a greatly reduced commitment of resources.

1 Cybersickness

From the earliest days of flight simulation, an attendant consequence of flying in visual flight simulators has been the occurrence of symptoms resembling motion sickness in crew members and instructors (Miller & Goodson, 1960). These symptoms include nausea, stomach awareness, sweating, disorientation, eyestrain, salivation, headache, and dizziness. Visually based symptoms (for example, eyestrain or dizziness) are slightly more prevalent in cybersickness than in other forms of motion sickness. Simulator sickness has been reported by persons training in all forms of visually based vehicular simulations (automobile, aircraft, tank) but has been most extensively documented in military flight simulators (Kennedy, Lilienthal, Berbaum, Baltzley, & McCauley, 1989).

Explosive technological progress in the last decade or so has powered an extension of virtual environment (VE) systems from research in the military to medical, educational, industrial, and entertainment applications (Burnett, 1998). This extension, however, has been accompanied by the same two problems that bedeviled the earlier development of simulator systems: potential cybersickness, and a possible transfer of maladaptive cognitive or psychomotor compensations from VE to real world environments with, as yet, unknown adverse legal, economic, individual, and social consequences (Stanney & Sal-

Presence, *Vol. 13, No. 5, October 2004,* 589–600

vendy, 1998). Developers of VE systems for entertainment have generally managed these issues by limiting exposure times to very short runs (2 to 4 minutes is typical). This solution, however, will not work for applications that require longer exposure times. Several reviews of this problem have appeared in recent years (Kennedy, Hettinger, & Lilienthal, 1990; Kolasinski, 1995; Stanney & Salvendy, 1998).

Biomedical applications of VE technology are especially promising (Machover, 1996; Mon-Williams & Pascal, 1995; Rosenberg, 1994). Current biomedical plans include training in surgical techniques, therapeutic applications for phobias, memory loss and rehabilitation, visualization of molecular structures, and many others. In demonstrations of such systems at trade shows, users are able to develop feelings of presence during their brief presentations. Experience with these compellingly realistic perceptions implies that training in such devices should transfer easily to real tasks. But here again there is an accompanying risk of cybersickness and associated side effects (e.g., disorientation and drowsiness). As exposure times in VE medical training lengthen to those comparable to actual surgical procedures, cybersickness symptoms are likely to become more nettlesome. In addition, recent studies with helmet-mounted VE devices have produced postexposure postural and eye-hand coordination disturbances (Kennedy & Stanney, 1997) that could interfere with normal postexposure activities such as driving and, if left unchecked, could also result in product liability claims.

2 Controlling Cybersickness

Controlling cybersickness, that is, holding it below a specified level for a large majority of users, is an eminently applied problem. It would help to have known empirical regularities on which one could depend. It would help too to have a good theory of cybersickness. Neither, however, necessarily suffices to control cybersickness in a specific device. Arcade operators (schools, military agencies) want to know if a new VE game (educational program, control station) will make people sick. To serve the purposes of control it is not enough that an empirical regularity hold on the average. It must hold in the particular device of interest. Further, it must be possible, given both the device and the regularity, to say just how much cybersickness the device will produce after a certain length of exposure. Theories of cybersickness must meet the same requirements. At this writing no regularity and no theory meets these requirements. In practice, therefore, arcade operators (schools and military agencies) must carry out experiments with the specific devices they propose to use.

This conclusion does not mean that empirical regularities and sound theory may not be helpful in the design of such experiments. Factors thought to contribute to cybersickness (determinants) fall into five groups. The first group consists of technical system factors such as optical distortion, field-of-view, flicker, motion platforms, refresh rate, resolution, transport delays, and update rate (Biocca, 1992; Kennedy, 1996; Kolasinski, 1995; Pausch, Crea, & Conway, 1992). The second group consists of user characteristics such as experience, gender, field independence, age, illness, critical flicker fusion threshold, mental rotation ability, postural instability, and susceptibility to motion sickness (Kolasinski, 1995; Kennedy, 1996; McCauley, 1984). The third group consists of a single determinant, namely, how long the subject remains in the virtual environment. The fourth group concerns the schedule of exposure, whether distributed over hours, days, or months. The fifth and final group consists of "kinematics," that is, all those variations in scene content and subject-system interactions that are affected by what the subject does during VE interaction, for example, position-tracking errors, turns, dives, altitude, and much else.

Length and distribution of exposure can be controlled directly. The only limitation is that a subject cannot be required to remain in a VE device if he or she is becoming ill. Technical system factors can also be controlled directly, although equipment design or redesign is often necessary. User characteristics can be controlled only indirectly, by selecting subjects who have specified properties or, in some instances, by giving different subjects different amounts of training or experience. The interactive nature of VE devices makes kinematics difficult but not impossible to control. Subjects could be

instructed to respond in specified ways and their compliance with instructions checked after the fact. Or, one could introduce a forcing function, for example, a constant bias in the left-right axis, that obliges a subject to exert continuous control to maintain a desired heading. To date, however, neither of these approaches has been attempted.

Many technical factors have been shown to affect cybersickness, some with respectable regularity and over several devices (Kennedy, 1996). Transport delays and refresh rates in particular have recently been confirmed as sickness determinants (Lackner & DiZio, 1998). The best established user characteristic is susceptibility itself. Individuals are measurably different in their vulnerability to motion sickness (Kennedy, Dunlap, & Fowlkes, 1990). Length and schedule of exposure are also well-established determinants, the latter mainly as adaptation (Kennedy, Stanney, & Dunlap, 2000; Welch, 2002). Nevertheless, these determinants, though numerous, are not strong and robust enough to allow containment of cybersickness below specified levels without trial in the device of interest.

It may be that all motion sickness regularities act through a final common path. It has been argued that sensory conflict is such a final common path (Reason & Brand, 1975). According to the conflict theory, all of the determinants listed above contribute to cybersickness by generating or facilitating conflict within or between sensory inputs. The sensory conflict theory is by no means universally accepted (Stoffregen & Riccio, 1991). Even if it were, while it might suggest hypotheses to be tested, it would not substitute for experimentation with the device of interest or concern. If the user wishes to configure a device so that only a few individuals experience more than a specified degree of cybersickness, generalities about the causes of motion sickness will not suffice.

Similar conclusions apply to other theories. Riccio and Stoffregen (1991, p. 205) have proposed that "prolonged postural instability is the cause of motion sickness symptoms." The same authors present evidence that in a fixed-based flight simulator "head motion among participants who later became sick was significantly greater than among participants who did not become sick" (Stoffregen, Hettinger, Haas, & Roe, 2000, p. 458). Control, however, implies causality. It could be that early head motion is a symptom of susceptibility. An equipment configuration that produced head motion might not make nonsusceptibile individuals sick. Conversely, an enforced absence of head motion (or postural instability) might not eliminate sickness in situations that would otherwise produce it. Stoffregen and Smart (1998, p. 446) have proposed such experiments—but until they are carried out and postural instability is demonstrated to be a robust causal determinant of motion sickness, their usefulness for purposes of control will be open to question.

This situation is not unusual in human-factors engineering. It often happens in applied work that there are no known empirical regularities or theories upon which one can rely for real-world manipulation and control. In all such situations, if control is to be achieved, it can only be by studying empirically how cybersickness or other "performance" of interest or concern varies as a function of its determinants in the device of interest. If these determinants were few in number, not more than four, say, such functional relationships could be worked out by means of conventional designs (complete factorials) without great difficulty. Unfortunately, the determinants are rarely, if ever, few in number.

3 Simon's Research Strategy

Given that the performance of interest has many manipulable determinants, the obvious first step is to screen those determinants with a view to finding out which ones are most important and which ones less so. If only a few factors are studied at a time, the inevitable result is a list of factors with little evidence as to their relative importance or the interactions between them, and no comprehensive account of how the performance of interest is functionally determined. It is somewhat surprising, therefore, that factorial studies involving more than a few factors are unusual in the human-factors literature. In a review of all experiments published in the journal *Human Factors* between 1958 and 1972 inclusive, Simon (1976b) counted 239 analysis-of-

variance tables. Of these 239 tables only 10.4% involved more than three, only 2.9% more than four, and only 0.4% more than five factors.

It is not, moreover, as if statisticians had not worked out the design of more complicated experiments. Fractional factorial designs and deliberate confounding date from the very beginning of analysis of variance (Fisher, 1935), were well advanced by the 1960s (Plackett & Burnam, 1946; Box & Hunter, 1961a, 1961b), and had been widely used in agricultural research, chemical engineering, and other applied fields. In the 1970s Charles W. Simon (1970a, 1970b, 1973, 1974, 1975, 1976a, 1976b, 1977a, 1977b) published a series of technical reports, together totalling more than 1,500 pages, in which he called attention to the literature on what he called "advanced experimental designs" and urged engineering psychologists to adopt them in their work. Apart from a series of studies utilizing the Visual Technology Research Simulator at the Naval Training Systems Center in Orlando, Florida (see below), his appeals went unheeded.

Simon assumes that the task of the engineering psychologist is to predict and control performance in real-world situations. He assumes that the performance of interest has many manipulable determinants of some importance, more than 5 and probably at least 15 or 20. Finally, he assumes that these determinants obey what he calls "Pareto maldistribution theory" (Juran, 1951). This theory states that although performance may be affected by many determinants, only a few are critical and many are trivial. Magnitudes of effect are distributed more or less like a chi-square distribution.

Simon contends that engineering psychology requires a program of research, not a miscellany of stand-alone experiments. It requires a series of experiments, most of which are not fixed from the beginning but may take somewhat different directions depending on the outcomes of experiments earlier in the series. This series, moreover, should be marked by what he calls *progressive iteration*.

As Simon uses the term, progressive iteration has three major implications. The first has already been mentioned, namely, that progressive iteration implies a *program* of research, not a collection of independent experiments. The second implication is that this program is *progressive*. It begins with a screening experiment, a rough "first cut," primarily designed to order known and suspected determinants of the performance of interest by magnitude of effect. Almost invariably this first cut will be a fractional factorial design in which all factors are represented by two levels. With such an experiment as a beginning, the program proceeds to locate and isolate two-way or even three-way interactions, where the basic experiment suggests they may have appreciable magnitudes of effect. Some of the more important quantitative factors may be nonlinear. So additional experiments (central composite designs) are performed to describe their response surfaces, alone or in combination with other factors.

The third implication of progressive iteration is *modularity*. Simon means by this term that experiments later in the series build on those that precede them. Each experiment is a block of treatment conditions. By itself, such a block may generate little or no information of value. It is informative only in combination with earlier blocks. Each new experiment is an extension of the experimental program as a whole. The original design, the screening experiment, allows for a large number of possible continuations. Only one of those continuations will, in fact, be realized. Nevertheless, that one was, in principle, considered in advance and provision made for it to be developed by extension from the initial experiment.

In developing his approach, Simon consistently emphasizes economy of design. No more treatment conditions, no more data points, should be included than are necessary to achieve the design objectives. These objectives, moreover, are not a matter of achieving statistical significance. Simon points out again and again that statistical significance is almost beside the point. The overall objective of his progressive iteration is to describe performance as a function of its determinants. The task is one of parameter estimation, not statistical reliability. Simon's insistence on this point has been much criticized (see below under "A Modified Simonian Approach").

Isoperformance curves and their utility in trade-off analysis were not formally developed (Jones & Kennedy,

1996; Jones, 2000) until 20 years after Simon wrote his technical reports. Their relevance in the present context is that the main concern with cybersickness is that it be held below a specified level a specified proportion of the time, and isoperformance is ideally suited to problems that take this general form, that is, that require the engineer or investigator to specify a level of performance to be reached or not to be reached, regardless of the determinantal combination used to do so.

Isoperformance analysis is based on the proposition that if one knows how performance varies as a function of its determinants, it is possible formally to derive equivalent combinations of the determinants, equivalent in the sense that all such combinations produce the same specified level of performance—hence the name *isoperformance*. What can't be accomplished in one way can usually be accomplished in other ways. The determinants trade off with one another. A little more of one determinant makes up for less of another. Which combination should be implemented is decided on the basis of nondeterminantal considerations, for example, cost, feasibility, side effects, safety, and the like.

Simon clearly understood the isoperformance approach and provided for it in his strategy. At one point, for example, he commented (1970b, p. 5) that one of the advantages of regression equations was that they could be used "to determine how equipment trade-offs should be made in order to optimize performance when one or more system parameters must be constrained." At another point (Simon, 1976a, p. 129), he noted that engineers ordinarily prefer information in the form of trade offs. "An engineer wants to know what will happen to performance if he [or she] uses a little less expensive component or if he [or she] improves one factor and degrades another, in order, for example, to reduce the weight or size of the equipment."

4 The VTRS Experiments

From 1979 to 1987 the Navy supported a large-scale study of simulator design for training purposes utilizing the Visual Technology Research Simulator (VTRS) at the Naval Training Systems Center in Orlando. The research was carried out under contract, initially by Canyon Research Group and then by Essex Corporation. Charles Simon was a member of the research team from the beginning of the project until its end, and his was the dominant point of view in the project's conception and design. Simon published an account of his approach in *Human Factors* (Simon & Roscoe, 1984) but the only review of the empirical work (Kennedy, Lane, & Fowlkes, 1989) is not available in the open literature.

The VTRS project was conducted in three phases. In the first phase, performance experiments were conducted in which experienced pilots were tested in the simulator under various experimental conditions. This type of experiment did not involve the transfer-of-training paradigm. However, it did serve as a vehicle for perfecting VTRS as a research tool and training device. In addition, the information obtained was useful in planning experiments at later stages. It was also directly relevant to the design of simulators for skill maintenance and transition training. This benefit was especially important as skill maintenance and transition training were substantially more expensive than undergraduate training (Orlansky & String, 1977a, 1977b). Finally, this first phase served as a screening device for experimental factors. Factors that showed little promise of meaningful training effect were dropped from further study at this point.

In the second phase, quasi-transfer experiments were conducted. Quasi-transfer experiments followed the transfer-of-training experimental paradigm but, after training in the simulator, testing also took place in the simulator under a standard high-fidelity configuration. This phase used novice pilots who were then trained and tested, so that results were directly relevant to the undergraduate pilot population. The quasi-transfer phase also served to screen variables and refine instructional methods.

In the third and final phase, training took place in the simulator and testing in the field, under a classic simulator-to-field transfer paradigm. This phase was, of course, definitive from a theoretical point of view but also extremely expensive and difficult to carry out logistically.

VTRS was a successful program in that the contribu-

tion of equipment factors to training was established. Negative results, however, were reported for most of the equipment features, indicating that higher fidelity was not required in most cases. Sufficient statistical power was, moreover, available to detect differences due to equipment variations if such differences existed. Regarding simulator motion, several studies compared motion-on vs. motion-off along with other equipment features. While other simulator configuration variables yielded intuitive, small, and sometimes statistically significant differences, motion was without significant effect.

A primary "lesson learned" from the VTRS research was the important contribution made by individual differences to performance. This result was evident in most of the experiments conducted at VTRS, even though individual differences were not generally an explicit part of the experimental design. From an equipment-oriented point of view, the VTRS results were disappointing. By and large, equipment and even instructional differences made much less difference than who the subject was.

5 A Modified Simonian Approach

Despite its rigor and sophistication, Simon's lead has not been followed. The VTRS program is the only exception. Two major reasons stand out for this neglect of Simon's work. The first is the size of the undertaking he recommends. The VTRS program lasted nine years and cost many millions of dollars. Even if the results had been more substantial than they were, not many investigators could seriously consider conducting (or securing funding for) so large an effort.

The second reason is Simon's attitude toward statistical significance. Most investigators today agree that significance alone is insufficient assurance that a study is nontrivial. Magnitude of effect must also be taken into account. Even fewer, however, would contend that magnitude of effect alone is adequate assurance of nontriviality. A small number of specific factors might indeed account for most of the variance but, absent statistical reliability, there can be no assurance that they will do so again, perhaps when a good deal depends on it.

Despite these weaknesses, Simon's approach has much to recommend it. He understood that important behavioral outcomes with many determinants posed a methodological problem. His invocation of "Pareto maldistribution theory" was appropriate. It is true that in most multiply determined outcomes a minority of factors account for most of the variance. He saw the need for screening experiments to identify which factors these were and he correctly pointed to fractional factorial experiments as a design strategy. Finally, he recognized the relevance of what would later be developed as "isoperformance methodology." The "modified Simonian approach" to be developed in this section is an attempt to implement the isoperformance logic in multiple determinant problems and to do so by incorporating many of Simon's ideas while at the same time avoiding those that led to his neglect.

As originally presented (Jones & Kennedy, 1996), isoperformance begins with a formal model of performance as a function of two or more determinants. The user (engineer or investigator) then specifies a level of performance he or she wishes to achieve and a probability that this level will be reached or exceeded. In a training situation, for example, the specified level of performance might be 70 (on a scale of 100) and the probability .90, that is, that 90% of the trainees graduate with a score of 70 or better. Simon's research strategy was aimed mainly at developing a functional model of performance. He recognized, however, that once he had such a model, isoperformance curves could be obtained by simply fixing performance at a specified level and solving. With two determinants, A and B, one could fix A (as well as the level of performance) and solve for B. With three determinants one could fix two of them and solve for the third. Any set of determinants obtained in this way suffices to produce the specified standards of performance (level and probability).

For control purposes, Simon's original program eventuated in a model-based isoperformance analysis. The modified Simonian approach differs in that it does not require a formal model. Isoperforming factor combinations are determined empirically, not derived analytically

from a model. One trains or exposes subjects until they reach specified standards of performance. In a learning context, for example, one might specify aptitude level and mode of instruction (the method used) and continue practice until 90% or more of the subjects reached a criterion of, say, 70. All combinations of aptitude level and mode of instruction that met these requirements in, say, 10 or fewer practice sessions (or the equivalent in instructional time) would be isoperforming. Which one of these combinations to adopt would then depend on nondeterminantal considerations, such as cost, administrative feasibility, the availability of students with the specified aptitudes, or the availability of appropriately trained instructors.

Several points are immediately in order. First, the empirical approach involves a change of dependent variable. In the model-based approach, the dependent variable is a measure of performance. In the case of cybersickness, for example, it might be the Simulator Sickness Questionnaire (Kennedy, Lane, Berbaum, & Lilienthal, 1993). In the empirical approach, the dependent variable is time, number of trials or sessions, or some other indicator of opportunity or exposure. Time is particularly apposite for cybersickness, because everyone agrees that the longer one remains in a sickness-provoking situation the more likely one is to become ill (Kennedy et al. 2000).

Second, the model-based approach is general: it applies wherever the model holds. If the model holds for all performance levels from 65 to 95, then any performance level within this range can be specified as not to be exceeded. The empirical approach is more limited. If the subject, after appropriate instruction and experience (see below), indicates when he or she has reached the specified level of sickness, then isoperforming curves or factor combinations can be obtained only for that level. Defensible curves might be obtained for the immediate neighborhood of the criterion level by linear extrapolation, but any level more than a short distance away would require a separate empirical determination.

Third, the empirical approach supposes some procedure for recognizing when a subject has reached the criterion level of sickness. If sickness is measured objectively, perhaps by electrogastrogram (Stern, Koch, Leibowitz, Shupert, & Stewart, 1985), then knowing when the criterion has been reached is a simple matter of observation. To find out how sickness-provoking a particular configuration of the video game (a particular combination of factors) is, subjects play the game in that configuration until they reach the criterion. The longer on average it takes the subjects to reach criterion, the less sickness-provoking that combination of factors is.

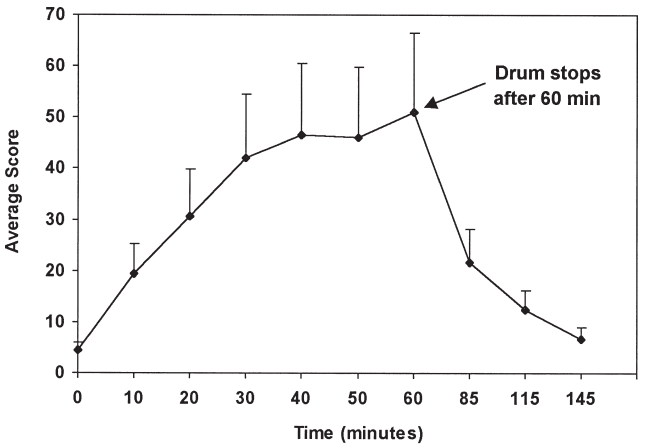

Figure 1. *Average score on an abbreviated version of the SSQ as a function of time in a rotating optokinetic drum. The subjects (N = 6) were tested every 10 minutes until the drum was stopped after one hour. The vertical bars indicate standard errors.*

If performance is measured by subjective report, the problem is more difficult. One possibility is to use a scale like the Simulator Sickness Questionnaire (SSQ) to identify the level of sickness not to be exceeded. The subjects would be instructed to recognize this level and given sufficient exposure to sickness-provoking situations, to experience it. To find out how sickness-provoking a specific configuration of the device is, subjects would be placed in the device so configured and remain there until they reached the criterion level they had been trained to recognize.

Figure 1 illustrates a second possibility. Six subjects were placed in a rotating optokinetic drum. Every 10 minutes the drum was stopped for approximately 4 minutes, while the subject was given an abbreviated version of the SSQ. The questions were asked and answered verbally. Then the drum was started up again for an-

other 10 minutes. After 60 minutes in the rotating drum, the subject was removed while SSQ testing continued at half-hour intervals.

The main point is that simulator sickness as measured by the SSQ rises with time very much as one would expect. One cannot be sure, of course, that the measured level is the same as it would have been if the drum had not been stopped for testing. It could be less and it could be more. It could also be just where it would have been if the subjects had experienced the same duration of uninterrupted exposure. The modified Simonian approach provides for a cross-validation in which inferences drawn from results like those depicted in Figure 1 can be checked.

At this point Simon's research strategy reenters the discussion. A single experiment is adequate to determine whether a particular combination of determinants qualifies as a possible option. The number of such combinations, however, may be very large. Suppose that each determinant (each factor) has only two levels, one that makes for sickness substantially more quickly than the other. With eight factors, a complete factorial would require 256 treatment conditions (separate experiments), many more than is feasible. Somehow the amount of experimental work needed to identify a respectable number of qualified options has to be reduced.

One way would be to identify on theoretical grounds a much smaller number of candidate combinations, say, 20, and test them. Of the 20 combinations, perhaps 7 would qualify as possible options. The remaining 13 would reach the specified standards of performance and probability in less time than is acceptable. If the 7 possible options were acceptable on other counts, such as cost, safety, and side effects, the user might be content.

Of course, matters might not work out so neatly. It might turn out that all of the combinations either resulted in cybersickness in less time than was acceptable or were unacceptable on other counts. Even if a few combinations survived all disqualifying considerations, there would be no empirical assurance that one or more of the combinations not tested might not have been preferable to any of those tested.

Screening designs, as described by Simon, offer a more systematic way of looking for possible options. In these designs, main effects are separated from each other but are confounded with two-factor and higher-order effects. The number of experimental conditions in such a design is a power of two and exceeds the number of factors by one. Thus, one might have, for example, 8 experimental conditions and 7 factors or 16 experimental conditions and 15 factors. We continue to assume that all factors have two levels.

Table 1. *Seven Factors and Their Treatment Levels*

		Treatment	
Factor	Descriptor	Low	High
A	Transport Delay	40 ms	80 ms
B	Susceptibility[a]	<1	>6
C	Roll	Off	On
D	Pitch	Off	On
E	Yaw	Off	On
F	Scene Content[b]	Simple	Complex
G	Sound	Off	On

[a]Defined in this example by scores on the Motion History Questionnaire, Personal Susceptibility Scale (Kennedy et al. 1990).

[b]Two different settings on the VE device and not responsive to the subject's behavior, therefore, not a kinematic variation.

Table 1 presents an illustrative list of 7 factors. The factor Sound is not expected to affect sickness. Screening designs assume that all two-factor and higher-order interactions are negligible. A good design does not take this assumption on faith. There should be checks. If main effect G turns out to be significant, then either the assumption that it does not affect sickness is false or one or more of the higher-order effects confounded with it is nonnegligible. If main effect G is a null effect, that is, the same treatment for both the Low and High levels, then the only explanation for a significant effect is nonnegligible confounded effects. A single check is only a hedge, of course. A high order interaction could be nonnegligible but not confounded with G, in which case it would not be detected. Still, some check is better than none.

Table 2. *Screening Design for Eight Treatment Groups with Seven Factors*

Grps	Treatments							A	B	C	D	E	F	G	I
1		b				f		−	+	−	−	−	+	−	+
2	a		c			f	g	+	−	+	−	−	+	+	+
3				d			g	−	−	−	+	−	−	+	+
4	a	b	c	d				+	+	+	+	−	−	−	+
5			c		e			−	−	+	−	+	−	−	+
6	a	b			e		g	+	+	−	−	+	−	+	+
7		b	c	d	e	f	g	−	+	+	+	+	+	+	+
8	a			d	e	f		+	−	−	+	+	+	−	+

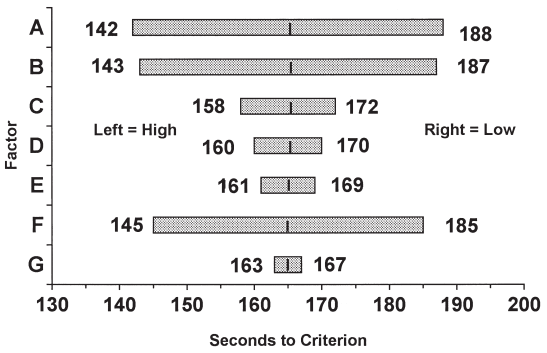

Figure 2. *Hypothetical results for a modified Simonian design. The vertical lines dividing the horizontal bars indicate the general mean at 165 seconds.*

Table 2 presents the design. The lowercase letters under the Treatments head refer to the High level of the corresponding main effect. Thus, all subjects in Group 1 experience the High level on factors B and F and the Low level on the other five factors. Subjects in Group 2 experience the High level on factors A, C, F, and G and the Low level on factors B, D, and E. The cell entries + and − to the right correspond to the lowercase letters under the Treatments head. The last column, "I," is the Identity column. All effects are contrasts and any two effects are orthogonal to each other.

Figure 2 presents a hypothetical outcome of the experiment in Tables 1 and 2. The bars indicate "effects" in the sense of the analysis of variance. That is, the bar for F, for example, indicates that the mean time to reach the criterion level of sickness is 20 seconds shorter for those who experience the Complex scene than for the average participant in the experiment. Similarly, time to criterion is 20 seconds longer for those who experience the Simple scene than for the average participant. The average time to criterion overall is 165 seconds, with an error variance within treatment conditions of 9 (standard deviation equal to 3). The example might be a video game. The designer wishes the game to be short enough that no more than 10% of those who play it reach the criterion he has indicated as just tolerable. He would also like it to be 180 seconds long. At least he would like to have the option of making it that long. Any combination of High and Low settings that meets these requirements will do.

Suppose, for example, that the designer sets Scene Content at High and all the others at Low. Mean time to criterion would then equal (165 − 20 + 23 + 22 + 7 + 5 + 4 + 2 =) 208. With an error standard deviation of 3, 90% would score at or above (208 − 3.84 ≅) 204, well above the desired game length of 180 seconds.

There is a problem, however. The designer cannot expect everyone who plays the game to be nonsusceptible. On average, they would be of middling susceptibility and, therefore, take close to the general average of 165 seconds to reach the criterion of just-tolerable sickness. Setting Scene Content at Complex, the other five equipment variations at Low, and letting Susceptibility

float, the average time to criterion drops to 186; and 90% would still not be worriesomely ill after 182 seconds.

There are still problems. Transport Delay, set at the Low condition, is likely to be expensive, perhaps prohibitively expensive. Further, the cutoff for 90% falls just 2.16 seconds above 180. There is bound to be sampling variation in the susceptibility of people who happen to play the game in a given period of time. In the Christmas season, the "regulars," who we may assume are less likely to be susceptible, may be likely to show off their pastime to friends and family who are perhaps more susceptible. The designer would then have to decide whether or not an adverse result over the holidays could be weathered.

No matter what the designer decides, a direct empirical check on his or her conclusions is necessary. A follow-up experiment could be conducted using the factor settings tentatively concluded as acceptable (for example, Scene Content set at Complex, the other five equipment variations Low, and Susceptibility distributed as it generally is). If one or more of the designer's requirements are not met, then either adjustments will have to be made (for example, by shortening the game or relaxing the 10% "error" rate), or the designer will have to consider other factor combinations. If all goes well, however, the designer will have arrived at a rationally defensible configuration with only the eight treatment conditions of Table 2 and one follow-up experiment. A full factorial experiment would have required 128 treatment conditions and at least one follow-up experiment.

The follow-up experiment is a cross-validation. It allows the experimenter to check on three potential sources of error. In selecting a combination of factor settings for follow-up testing the designer capitalizes on chance. When tested in an independent sample, the results will probably not be so favorable. Depending on just how much "shrinkage" occurs, what seemed to be an acceptable combination of factors may turn out not to be.

The main experiment did not take statistical reliability into account. In the follow-up experiment, significance becomes a major issue. Sample size should be at least as large as the total number of subjects in the screening experiment. The tolerable error rate of 10% is based on estimates of the mean time to criterion and the error standard deviation, both of which are subject to sampling error; and the only adequate protection is sample size. The proportion of subjects who remain in the follow-up experimental situation 180 seconds without reaching the criterion of just-tolerable illness is an empirical result. It is possible, however, to calculate how this proportion would be distributed if the experiment were repeated many times. The larger the sample, the greater its statistical reliability, and the better the probability that a positive result will be repeated.

Finally, if the method illustrated in Figure 1 is used for measuring time to criterion, there is still another possible source of difference between the time to criterion calculated for a specific combination of factors and the time observed in the follow-up experiment. With experience using this method, it should become clear whether the likelihood is to over- or underestimate time to criterion, and an appropriate adjustment can be made. Even so, however, a check should be made.

The modified Simonian approach is a method. Its purpose, moreover, is not primarily investigative but applied, namely, to configure variations in the equipment (transport delays, other display asynchronies, scene content) so as to hold cybersickness at or below acceptable levels. Cybersickness relates to perception and action in virtual environments primarily as an unwanted response. Nevertheless, the effort to control this response may well prove to be informative about the conflicting processes that give rise to it.

Acknowledgment

The authors thank Dr. Jonathan French for permission to present the data in Figure 1.

References

Biocca, F. (1992). Will simulator sickness slow down the diffusion of virtual environment technology? *Presence: Teleoperators and Virtual Environments, 1,* 334–343.

Box, G. E. P., & Hunter, J. S. (1961a). The 2^{k-p} fractional factorial designs. Part I. *Technometrics, 3,* 311–351.

———. (1961b). The 2^{k-p} fractional factorial designs. Part II. *Technometrics, 3,* 449–458.

Burnett, R. (1998, December 11). Simulation belt may expand. *The Orlando Sentinel,* pp. B1, B6.

Fisher, R. A. (1935). *The design of experiments* (1971 reprint). New York: Hafner.

Jones, M. B. (2000). Isoperformance and personnel decisions. *Human Factors, 42,* 299–317.

Jones, M. B., & Kennedy, R. S. (1996). Isoperformance curves in applied psychology. *Human Factors, 38,* 167–182.

Juran, J. (1951). *Quality-control handbook.* New York: McGraw-Hill.

Kennedy, R. S. (1996). *Analysis of simulator sickness data* (Technical Report under Contract No. N61339-91-D-0004 with Enzian Technology, Inc.). Orlando, FL: Naval Air Warfare Center, Training Systems Division.

Kennedy, R. S., Dunlap, W. P., & Fowlkes, J. E. (1990). Prediction of motion sickness susceptibility. In G. H. Crampton (Ed.), *Motion and space sickness* (pp. 179–215). Boca Raton, FL: CRC Press.

Kennedy, R. S., Lane, N. E., & Fowlkes, J. E. (1989). *Review of human performance research at the Visual Technology Research Simulator* (Technical Report NTSC TR89-020). Orlando, FL: U.S. Naval Training Systems Center.

Kennedy, R. S., Hettinger, L. J., & Lilienthal, M. G. (1990). Simulator sickness. In G. H. Crampton (Ed.), *Motion and space sickness* (pp. 317–341). Boca Raton, FL: CRC Press.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *International Journal of Aviation Psychology, 3*(3), 203–220.

Kennedy, R. S., Lilienthal, M. G., Berbaum, K. S., Baltzley, D. R., & McCauley, M. E. (1989). Simulator sickness in U.S. Navy flight simulators. *Aviation, Space, and Environmental Medicine, 60,* 10–16.

Kennedy, R. S., & Stanney, K. M. (1997). Aftereffects of virtual environment exposure: Psychometric issues. In M. J. Smith, G. Salvendy, & R. J. Koubek (Eds.), *Design of computing systems: Social and ergonomic considerations* (pp. 897–900). Amsterdam: Elsevier.

Kennedy, R. S., Stanney, K. M., & Dunlap, W. P. (2000). Duration and exposure to virtual environments: Sickness curves during and across sessions. *Presence: Teleoperators and Virtual Environments, 9,* 463–477.

Kolasinski, G. (1995). *Simulator sickness in virtual environments* (Technical Report 1027). Orlando, FL: United States Army Research Institute for the Behavioral and Social Sciences.

Lackner, J. R., & DiZio, P. (1998). Personal communication.

Machover, C. (1996). What virtual reality needs. *Information Display, 12*(6), 32–34.

McCauley, M. E. (Ed.). (1984). *Research issues in simulator sickness: Proceedings of a workshop.* Washington, DC: National Academy Press.

Miller, J. W., & Goodson, J. E. (1960). Motion sickness in a helicopter simulator. *Aerospace Medicine, 31,* 204–212.

Mon-Williams, M. A., & Pascal, E. (1995). Virtual reality displays: Implications for optometrists. *Optometry Today, 35,* 30–33.

Orlansky, J., & String, J. (1977a). *Cost-effectiveness of flight simulators for military training: Volume I. Use and effectiveness of flight simulators* (IDA Paper P-1275). Arlington, VA: Institute for Defense Analyses.

———. (1977b). *Cost-effectiveness of flight simulators for military training: Volume II. Estimating costs of training in simulators and aircraft* (IDA Paper P-1275). Arlington, VA: Institute for Defense Analyses.

Pausch, R., Crea, T., & Conway, M. (1992). A literature survey for virtual environments: Military flight simulator visual systems and simulator sickness. *Presence: Teleoperators and Virtual Environments, 1*(3), 344–363

Plackett, R. L., & Burnam, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika, 33,* 305–324.

Reason, J. T., & Brand, J. J. (1975). *Motion sickness.* New York: Academic.

Riccio, G. E., & Stoffregen, T. A. (1991). An ecological theory of motion sickness and postural instability. *Ecological Psychology, 3,* 195–240.

Rosenberg, L. B. (1994). Medical applications of virtual reality. *Virtual Reality Systems: Applications Research & Development, 1*(3), 48–50.

Simon, C. W. (1970a). *Reducing irrelevant variance through the use of blocked experimental designs* (Technical Report No. AFOSR-70-5). Culver City, CA: Hughes Aircraft Company.

———. (1970b). *The use of central-composite designs in human factors engineering experiments* (Technical Report No. AFOSR-70-6). Culver City, CA: Hughes Aircraft Company.

———. (1973). *Economical multifactor designs for human factors engineering experiments* (Technical Report No. P73-326A). Culver City, CA: Hughes Aircraft Company.

———. (1974). *Methods for handling sequence effects in human factors engineering experiments* (Technical Report No. P74-541A). Culver City, CA: Hughes Aircraft Company.

———. (1975). *Methods for improving information from "undesigned" human factors experiments* (Technical Report No. P75-287). Culver City, CA: Hughes Aircraft Company.

———. (1976a). *Response surface methodology revisited: A commentary on research strategy* (Technical Report No. CWS-01-76). Westlake Village, CA: Canyon Research Group.

———. (1976b). *Analysis of human factors engineering experiments: Characteristics, results, and applications* (Technical Report No. CWS-02-76). Westlake Village, CA: Canyon Research Group.

———. (1977a). *Design, analysis, and interpretation of screening designs for human factors engineering research* (Technical Report No. CWS-03-77A). Westlake Village, CA: Canyon Research Group.

———. (1977b). *New research paradigm for applied experimental psychology: A system approach* (Technical Report No. CWS-04-77A). Westlake Village, CA: Canyon Research Group.

Simon, C. W., & Roscoe, S. N. (1984). Application of a multifactor approach to transfer of training research. *Human Factors, 26,* 591–612.

Stanney, K. M., & Salvendy, G. (1998). Aftereffects and sense of presence in virtual environments: Formulation of a research and development agenda. *International Journal of Human-Computer Interaction, 10*(2), 135–187.

Stern, R. M., Koch, R. L., Leibowitz, H. W., Shupert, C. L., & Stewart, W. R. (1985). Tachygastria and motion sickness. *Aviation, Space, and Environmental Medicine, 56,* 1074–1077.

Stoffregen, T. A., Hettinger, L. J., Haas, M. W., & Roe, M. M. (2000). Postural instability and motion sickness in a fixed-base flight simulator. *Human Factors, 42,* 458–469.

Stoffregen, T. A., & Riccio, G. E. (1991). An ecological critique of the sensory conflict theory of motion sickness. *Ecological Psychology, 3,* 159–194.

Stoffregen, T. A., & Smart, L. J., Jr. (1998). Postural instability precedes motion sickness. *Brain Research Bulletin, 47,* 437–448.

Welch, R. B. (2002). Adapting to virtual environments. In K. M. Stanney (Ed.), *Handbook of virtual environments: Design, implementation, and applications* (pp. 619–636). Mahwah, NJ: Erlbaum.