Behavior Research Methods
2006, 38 (2), 314-324

Using self-reported data to assess the validity of driving simulation data

BRYAN REIMER
Massachusetts Institute of Technology, Cambridge, Massachusetts
and Massachusetts General Hospital, Boston, Massachusetts

LISA A. D'AMBROSIO, JOSEPH F. COUGHLIN, and MICHAEL E. KAFRISSEN
Massachusetts Institute of Technology, Cambridge, Massachusetts

and

JOSEPH BIEDERMAN
Massachusetts Institute of Technology, Cambridge, Massachusetts
Massachusetts General Hospital, Boston, Massachusetts
and Harvard Medical School, Boston, Massachusetts

In this article, we use self-reported driving behaviors from a written questionnaire to assess the measurement validity of data derived from a driving simulation. The issue of validity concerns the extent to which measures from the experimental context map onto constructs of interest. Following a description of the experimental methods and setting, an argument for the face validity of the data is advanced. Convergent validity was assessed by regressing behaviors observed in the driving simulator on self-reported measures of driving behaviors. Significant relationships were found across six measures: accidents, speeding, velocity, passing, weaving between traffic, and behavior at stop signs. Concurrent validity was evaluated with an analysis of simulator accident involvement and attention deficit hyperactivity disorder status. Discriminant validity was assessed using a multitrait–multimethod matrix of simulator and questionnaire data. We concluded that although the relationship between self-reported behaviors and observed responses in the simulator falls short of perfect correspondence, the data collected from the driving simulator are valid measures of the behaviors of interest.

Researchers have identified medical disabilities, chronic disease conditions, and pharmacological interventions as factors that can influence people's physical and cognitive abilities to drive safely (Holland, Handley, & Feetam, 2003; Rimm & Hakamies-Blomqvist, 2002; Stutts & Wilkins, 2003). The National Transportation Safety Board (NTSB) recently noted its concern about the impact of medical conditions in the noncommercial driving population on safety (National Transportation Safety Board, 2004). The NTSB identified a "need for more data on the extent to which medical conditions contribute to the cause of accidents." Surveys have found an increased probability of driving accidents among those with conditions ranging from, for example, attention deficit hyperactivity disorder (ADHD; Barkley, Guevremont, Anastopoulos, DuPaul, & Shelton, 1993; Murphy & Barkley, 1996), to diabetes (Cox et al., 2003), to dementia (Carr, Duchek, & Morris, 2000; Zuin, Ortiz, Boromei, & Lopez, 2002), to heart disease (McGwin, Sims, Pulley, & Roseman, 2000), to arthritis (McGwin et al., 2000), and to emotional stress (Lagarde et al., 2004). Yet studies with such populations using on-road methods to measure driving behaviors may place both participants and researchers at increased risk for harm (Cox, Humphrey, Merkel, Penberthy, & Kovatchev, 2004). On-road behavior evaluations also lack the replicability, control, efficiency, safety, and ease of use associated with simulated driving experiments, in turn making inferences about the impact of experimental manipulations on driving behavior more difficult (Godley, Triggs, & Fildes, 2002). Even the most advanced simulators, however, lack the same physiological stimulation experienced while driving a real vehicle, often resulting in some degree of simulator sickness among some research participants (Ranney et al., 2002). Although critics rightly argue that people may react differently in driving simulators since there is no risk of collision or physical harm, laboratory environments provide the best alternative for addressing questions that are too dangerous or expensive to answer in on-road evaluations (Alm & Nilsson, 1995; Hahn & Tetlock, 1999; Hoffman, Lee, Brown, & McGehee, 2002).

This project was supported by the Johnson & Johnson Center for Pediatric Psychopathology Research at the Massachusetts General Hospital and the U.S. Department of Transportation's New England University Center at the Massachusetts Institute of Technology. Correspondence should be addressed to B. Reimer, Massachusetts Institute of Technology, AgeLab, 77 Massachusetts Ave., Room E40-291, Cambridge, MA 02139 (e-mail: reimer@mit.edu).

Driving simulators are merely tools for collecting data, albeit expensive ones. To draw inferences confidently about real driving behaviors from driving simulation data requires establishing the validity of the simulation. In short, do the behaviors observed in increasingly complex experimental simulations map onto drivers' behaviors in real-world driving experiences? The issue is one of measurement validity: To what extent does the simulator produce driving behaviors that are comparable to real driving behaviors? The issue of validity is particularly acute as researchers seek to understand the impact of various medical conditions and interventions on driving behaviors (e.g., for ADHD, see Cox, Merkel, Penberthy, Kovatchev, & Hankin, 2004; for schizophrenia, see Brunnauer, Laux, Geiger, & Möller, 2004; for Alzheimer's disease, see Rizzo, Reinach, McGehee, & Dawson, 1997). The gold standard of validating simulation protocols has typically been through a comparison of simulation data with data collected from on-road driving; however, in this study, we take a different approach. After providing a definitional framework to use for assessing validity, we explore the prospects for using survey data on self-reported driving behaviors to validate simulator measures.

Simulation Validation in Past Research on Driving Behavior

Simulation provides a cost-effective and efficient way to train people or measure performance on tasks that would be too dangerous, difficult, or expensive to do in the real world. For example, the field of medicine has increasingly used simulation to train medical students in behaviors as diverse as patient management, clinical diagnostic skills, and surgery (Berg et al., 2001; Dawson & Kaufman, 1998; Hawkins et al., 2004). For the same reasons of cost and risk, aviation flight simulators have been frequently used to study pilot behavior (Döring, 1990), to conduct pilot training (Rolfe & Hampson, 2003), and to evaluate the design of aircraft (Sarathy & Higman, 1994). Although flight simulators are widely employed for pilot training, where important decisions on crew certification are often made, there is "no substantial body of knowledge to predict or measure the effectiveness of flight training devices" (Rolfe & Hampson, 2003). Similarly, human factors research has long used on-road and simulated driving tasks to assess driver behavior and performance in the presence of a variety of technological innovations (Ben-Yaacov, Maltz, & Shinar, 2002; Brown, Tickner, & Simmonds, 1969; Decina, Gish, Staplin, & Kirchner, 1996; Landau, Laur, Hein, Srinivasan, & Jovanis, 1994; McKnight & McKnight, 1993; Nunes & Recarte, 2002; Sodhi, Reimer, & Llamazares, 2002). There are two problems with the validity of measures from many driving simulation studies. First, few studies address the issue of validity at all, instead ignoring or assuming it (for a list of studies that do address validation directly, see Godley et al., 2002). Second, there is a lack of consensus on the vocabulary used to describe and establish validity. While social scientists have distinguished among measurement, internal, and external validity, much of the driving simulation literature has instead collapsed the discussion of measurement validity into what is labeled *behavioral validity* of two types—absolute and relative (Godley et al., 2002; Kaptein, Theeuwes, & van der Horst, 1996; Törnros, 1998).

The absolute behavioral validity of an indicator from a driving simulation study is its perfect correlation with measurements of the same behavior in real-world or test-track driving (Godley et al., 2002; Harms, 1996; Kaptein et al., 1996). The relative behavioral validity of an indicator does not require its perfect correlation with a measure from real driving behaviors, but it does require that the two different measures be in the same direction (Godley et al., 2002; Harms, 1996; Kaptein et al., 1996). For example, some researchers have demonstrated the relative validity of measures, such as speed control (Reed & Green, 1999), dialing simulated cellular telephones (Blana & Golias, 2002), and lateral displacement on straight and curved roads (H. C. Lee, Cameron, & A. H. Lee, 2003), through a comparison of real-world driving experiments and comparable experimental laboratory simulations.

Behavioral validity in essence requires the development and comparison of data from concurrent real-world and simulated driving studies. Such methods of assessing validity are costly and often impractical. Thus, questions about the validity of driving simulation data are most frequently left unasked and unanswered. Exceptions to this include the studies of Klee, Bauer, Radwan, and Al-Deek (1999) and Stanton, Young, Walker, Turner, and Randle (2001), research in which the authors compared measures of on-road driving with simulated driving behaviors that used simulation images designed to mimic the particular on-road conditions from their studies. Klee et al. (1999) compared forward speed, whereas Stanton et al. (2001) used measures from a secondary task to demonstrate the similarity of cognitive workload levels and psychological environments in on-road and simulated conditions. The rapid prototyping of simulation scenarios (Allen, Park, Rosenthal, & Aponso, 2004), however, has led to the development of increasingly complex driving simulation scenarios. At issue is whether data from these types of simulation scenarios are valid indicators of the real-world behaviors of interest to researchers.

In this article, we focus on establishing the measurement validity of indicators from a long-duration driving simulation scenario. We begin by defining and describing different facets and types of validity common in social science research and by mapping these terms for validity onto those used previously by other researchers. We then look at a variety of different measures from a simulation scenario in order to establish their validity by comparing them with self-reported survey questionnaire data. Such a method of validation has not been used previously in driving simulation studies, although it is more common in other social science disciplines (e.g., Rosenstone, Hansen, & Kinder, 1986). Should this method prove successful, it would provide a faster and less expensive means for other researchers to establish the validity of measures collected from different driving simulation scenarios.

DEFINING VALIDITY

Social scientists typically identify three types of validity: measurement, internal, and external. Measurement validity concerns whether an indicator actually measures the concept of interest. Internal validity and external validity relate to researchers' abilities to draw inferences from the results of the data collection. The latter two types of validity are primarily affected by the overall nature of the experimental design rather than any individual measure. Although others (e.g., Godley et al., 2002; Kaptein et al., 1996; Reed & Green, 1999) have used different language to discuss validity in simulation research, these conceptions of validity fall within the three broader categories. The following sections present an overview and definitions of different concepts of validity related to driving simulation research. Table 1 presents a summary and definitions of the different types of validity.

Measurement Validity

"The concern with which construct an instrument measures is the concern about its validity" (Kidder & Judd, 1986, p. 50). In short, to what extent do the measures of interest accurately tap or capture the underlying concept of interest? In the case of the research we discuss here, the question is the extent to which driving behaviors observed in simulation experiments are indicative of people's driving behaviors in the real world. Four aspects of validity are typically considered when assessing a measure's validity: face validity, concurrent validity, predictive validity, and discriminant validity. The first three types of validity are different facets of convergent validity, the degree to which one means of measuring a construct agrees with another (Kidder & Judd, 1986). Establishing these different types of validity for a measure increases overall confidence that the indicator measures the concept it is intended to. Table 1 provides a map of the different types of validity and their relationship with the vocabulary used in previous simulation research.

Convergent Validity

Face validity. Face validity is the degree to which a group of experts agree that in their opinion the measure captures the intended construct. It is most closely aligned with what others have labeled *physical validity* (e.g., Godley et al., 2002). The physical validity of a driving simulator correlates with its degree of realism. Descriptions of the physical or face validity provide technological details about the simulator, the simulation, and the setting that in principle mimic real-life driving more closely. Although no driving simulator can wholly replicate the actual experience of driving, high-fidelity simulators with motion platforms presumably offer the highest level of face or physical validity (Godley et al., 2002). Nevertheless, far less complex systems are capable of providing comparable or better validated measures of driving behavior (J. D. Lee, 2004; Reed & Green, 1999).

Concurrent validity. "Concurrent validity is the ability of a test to distinguish between individuals who are known to differ" (Kidder & Judd, 1986, p. 55). For measures with high concurrent validity, people known to differ on another measure related to the indicator of interest should score differently on the basis of the group they belong to. For example, we might expect that a measure with high concurrent validity from driving simulator data, such as average speed in the simulation, would be related to the number of real-life speeding tickets people reported.

Predictive validity. "Predictive validity is the ability of a test to identify future differences" (Kidder & Judd, 1986, p. 55). A frequent example of predictive validity revolves around the use of Scholastic Aptitude Test (SAT) scores as an indicator of success in college. If SAT scores are a valid indicator of this construct, then we should expect that SAT scores should predict students' future college performance, such as their grade point averages. In the case of driving simulation, predictive validity means that measures of driving performance from the simulator data would predict real-life driving performance.

Discriminant Validity

Discriminant validity is the extent to which a construct "can be differentiated from other constructs, and to dem-

Table 1
Definitions of Validity

Type	Definition	Other Terms Used to Describe Validity
Measurement validity	The extent to which an indicator measures the construct or concept of interest	Construct validity
Convergent validity	Means of assessing validity that depend on agreement	Absolute behavioral validity; relative behavioral validity
—Face validity	The extent to which experts agree that the measure captures the intended construct	Fidelity; physical validity
—Concurrent validity	Whether the measure can distinguish between individuals or groups known to be different in some way relative to the measure	
—Predictive validity	Whether a measure can be used to identify future differences	
Discriminant validity	Differentiation of the construct from measures of other, different constructs	Construct validity
Internal validity	The extent to which causal inferences about the impact of an experimental treatment can be made confidently	
External validity	The ability to generalize experimental results to other populations, time periods, or settings	

onstrate this a researcher must show disagreement between two scores that presumably measure different constructs" (Kidder & Judd, 1986, p. 56). In driving simulation research, for example, we would expect there to be little agreement between measures of driving performance and personal time management skills.

Internal Validity

Internal validity is the degree to which causal inferences about the impact of an experimental manipulation can be made confidently. Experimental designs are internally valid to the extent that any differences observed in participants' behavior can be attributed directly to the experimental manipulation and to no other causes (Campbell & Stanley, 1963). The degree of control over the simulation environment means that internal validity tends to be greater in driving simulation studies rather than in real-life driving experimentation. Threats to internal validity are factors that could confound with the experimental manipulation in order to produce the observed effects. These factors include history, maturation, selection bias, experimental mortality, instrumentation, testing effects, selection–maturation interaction, and regression effects (Campbell & Stanley, 1963).

External Validity

External validity is to the ability to generalize results obtained through a set of experiments to other populations, time periods, or environments (Kaptein et al., 1996). Campbell and Stanley (1963, p. 17) note that "the problems of external validity are not logically solvable in any neat, conclusive way. Generalization always turns out to involve extrapolation into a realm not represented in one's sample." External validity rests on the assumptions and arguments that the relationship between the experimental stimulus and behaviors in the laboratory setting is the same in the real world. Threats to external validity are those factors or interactions between variables in the experimental setting that impede or limit researchers' abilities to reproduce such effects in the real world. These include interaction of the experimental treatment with selection bias, with maturation, with history, or with testing, multiple treatment interference, and reactive arrangements (Campbell & Stanley, 1963; McGaw & Watson, 1976). The limited scope of most driving simulations, as well as the experimental setting itself, reduces their degree of external validity. Nevertheless, by designing simulation studies with the "maximum similarity of experiments to the conditions of application which is compatible with internal validity," researchers can attempt to maximize external validity (Campbell & Stanley, 1963, p. 18).

METHOD

Participants

A total of 48 active drivers were selected to participate in a pilot driving simulation study; 25 participants were known to have ADHD. ADHD is distinguished by impairments in attention and/or impulse control, contributing to significant difficulties in academic, social, and occupational functions (Biederman & Faraone, 2004). All ADHD participants met full DSM-IV criteria, had symptom onset in childhood, and had persistent symptomatology into adulthood. The participants were classified as controls if they failed to meet the diagnostic criteria for ADHD or had fewer than three ADHD symptoms. The participants were between 16 and 55 years of age and spoke English. Excluded were those with IQ scores of less than 80 and those with other DSM-IV diagnoses. The participants were recruited through clinical referrals to an adult ADHD program at a major medical center and through advertisements in the local media. All participants were required to sign two consent forms approved by the institutional review boards from each institute. The average age of the participants was 30.41 years (*SD* = 8.49, range 18–51 years). Twenty-three of the participants were male.

Procedure

Eligible participants completed four written questionnaires before and one following simulation testing: a U.S. variation on the U.K. Driver Behavior Questionnaire (Lawton, Parker, Manstead, & Stradling, 1997; Parker, Reason, Manstead, & Stradling, 1995; Stradling & Meadows, 2000), a survey that collected information about each participant's driving history, a health information questionnaire, and a presimulator and postsimulator sickness survey. Once the participants were comfortably seated in the simulator, a recorded series of instructions was repeated in accordance with the informed consent procedures of both institutions. Recorded instructions included details on the training and testing procedures, instructions on the operation of a hands-free cellular telephone, and instructions on a visual display response system for cognitive tasks during the simulation. The participants also heard details of performance-based incentives and the procedures they should follow in the case of any discomfort or "simulator sickness." After the participants received the recorded instructions, a researcher answered any final questions. Just prior to the beginning of the simulation, the participants were reminded to inform the operator immediately in the event of any discomfort. In the case of reported discomfort likely linked to simulator sickness, the participants were advised to stop driving if there were any further increase in symptoms.

Scenario

The driving environment included virtual training designed to last about 10 min, followed by a longer high-stimulus urban testing segment (Segment 1), and concluded with a longer low-stimulus rural testing segment (Segment 2). There was a slight pause between each segment. The training allowed the participants to become accustomed to the controls of the car and other features to be presented in the testing segments, including traffic control devices (traffic lights and stop signs), other vehicles, road grades (up and down), lane configurations (single and multiple), and curves. The two testing segments were designed such that a driver, following the speed limit, would require approximately 45 min of driving time to complete both segments. Both segments are described in more detail below. A complete overview of the scenario is presented in Figure 1.

Segment 1. Figure 1 shows that six components made up the urban testing segment. The components were divided into three conditions: suburban, urban, and car following. In the suburban condition, the speed limit was 45 mph; in the urban and car-following conditions, it was 35 mph. In each "repeated" condition (i.e., suburban and urban), events were repeated in a different order, although there were no differences in the scenario between drivers. In the suburban condition, drivers encountered a series of five traffic lights (three of which changed color as the driver approached). Each suburban condition contained four events, a jaywalker and cars backing suddenly out of a driveway, each presented once with and once without a

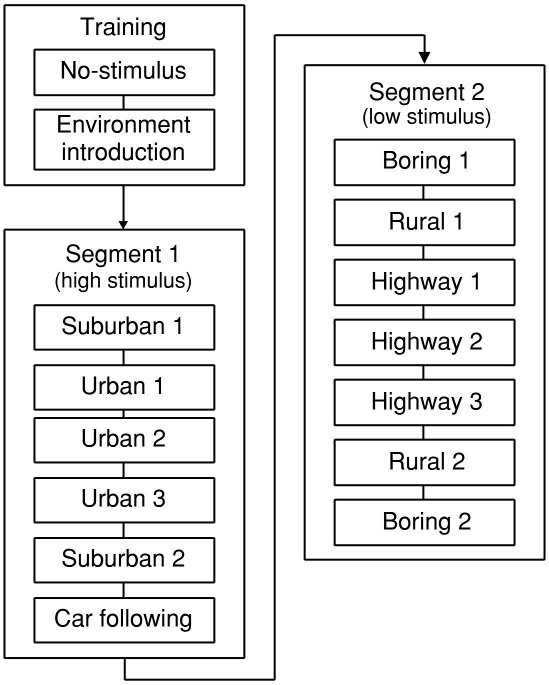

Figure 1. Flow of the scenario presented in the driving simulator.

lead vehicle. The urban condition focused on the use of stop signs to force the driver to alternate between acceleration and deceleration. Events included a jaywalker, a parked car pulling slowly in and out of the lane of travel, and pedestrians in a crosswalk. The car-following condition introduced low-stimulus driving with oncoming traffic to constrain the driver to follow a lead vehicle with a velocity changing according to a sinusoidal function with mean 35 mph.

Segment 2. Three different conditions across seven components made up the low-stimulus testing segment detailed in Figure 1. The speed limits were 55 mph in the rural and boring (no stimulus) environments and 65 mph in the highway. The rural conditions incorporated terrain modulation—hills and curves, while the highway presented two travel lanes as part of a four-lane highway. Events in the rural condition included passing a slow vehicle, reacting to an oncoming vehicle wandering into the travel lane, and a speed trap. Two types of events occurred on the highway. First, the driver approached a set of vehicles traveling at 50 mph in the right lane. When passing on the left, one of the cars quickly pulled in front of the driver without speeding up. The second event involved a series of vehicles moving at 50 mph in both travel lanes. To maintain an acceptable travel speed the driver was forced to weave through the traffic and thus to pass on the right. The final event presented in both of the boring conditions involved two dogs that appeared just off to the side of the road. When the driver reached a predetermined time to collision, the dogs began to run toward the road. In order to avoid an accident, drivers had to take some evasive action.

Incentives for Participation

Critics of simulation argue that, without perceived risk, drivers will respond differently in the experimental setting than they would in the real world. To compensate for this limitation, as Stein, Allen, and Parseghian (1992) did, we provided financial incentives to encourage people to maintain speed, obey the traffic laws, and devote attention to secondary cognitive tasks. The participants were told that they would receive $40 for their participation but could earn an additional incentive of $20 in each of three performance areas. They were instructed that they needed to complete the simulation in 45 min and would be penalized $1 per minute for each minute over 45 that it took them to complete the simulation. To promote safe driving, the participants were "charged" $5 for each crash and $1 for each ticket. Finally, they were told that they could increase their compensation by correctly answering questions involved with secondary tasks.

Survey Instrument

The driving history questionnaire included questions on basic driving history, current driving habits, accident and traffic violation history, and attitudes toward driving. A subset of the questions were adapted from the *National Survey of Speeding and Unsafe Driving Actions* (Boyle, Dienstfrey, & Sothoron, 1998). Table 2 presents a summary of the items used to validate the driving behavior simulation measures presented in Table 3.

RESULTS

Forty-one of the 48 participants who began the simulation portion of the experiment completed the driving task. All seven participants who failed to complete the experiment reported experiencing simulator sickness within 10 min of beginning the experimental portion of the protocol. The participants who reported simulator sickness were advised that continuing with the experiment would likely increase the severity of symptoms. Of these 7 par-

Table 2
Survey Items Used to Establish Validity

Variable Name	Question
Rate from 1 to 5 the most recent time you did the following, based upon the following scale: (1) Today (2) Within the past week (3) Within the past month (4) More than a month ago (5) Not in the past year	
Drove in traffic switching between lanes	Drove through traffic by switching quickly back and forth between lanes?
Failed to slow at a stop sign	Drove through a stop sign without slowing?
Slowed but failed to stop at a stop sign	Slowed but did not completely stop at a stop sign?
Pass in a no-passing zone	Passed a vehicle in a no-passing zone?
How many times while driving have you been pulled over and cited or warned (verbally or written) for:	
Total speeding tickets	Speeding more than 20 mph over the limit? Speeding more than 10 mph but less than 20 mph over the limit? Speeding, but for less than 10 mph over the limit?
Five-year accident history	How many times has this [being in a vehicle crash as either a passenger or driver] happened to you (in the past five years)?

Table 3
Simulator Measures Used to Establish Validity

Measure Name	Description
Traffic weaving	Mean time required on a four-lane highway (speed limit 65 mph) to pass three sets of cars traveling at 50 mph; one set of cars is in the left lane and two in the right (seconds).
Speeding	Number of 25-foot intervals where the driver's average velocity exceeded the speed limit.
Pause time at stop sign	Mean pause time for stop signs (seconds)
Lane variability	Number of intervals where the driver was one standard deviation to the left of his/her average lane position.
Total accidents	Total number of accidents during the simulation.
Mean control velocity	Mean velocity during a control period (normalized to the speed limit).

ticipants, all but one heeded the advice and withdrew from the study.

Establishing Convergent and Discriminant Validity of the Driving Simulation Measures

In this work, we focus specifically on establishing the measurement validity associated with measures from the driving simulation scenario. We examine two types of convergent validity, face and concurrent, as well as discriminant validity. The discussion of face validity provides an argument about facets of the experimental setting that increase the fidelity of the simulated driving experience. An analysis of collected survey and simulator data are used to develop arguments around concurrent validity. Finally, we construct a multitrait–multimethod matrix to examine discriminant validity.

Face validity. The driving simulator used includes a full 2001 Volkswagen Beetle cab. The software, STISIM Drive version 2.03.01, receives inputs from the original equipment manufacture (OEM) accelerator, brake, and steering wheel. Feedback to the driver is provided through visual, auditory, and kinetic channels. The visual representation of the roadway, updated at 20 Hz, is displayed on an 8 × 8 in. screen at a resolution of 1,024 × 768 pixels. This creates approximately a 40º field of view. The rearview mirror in the OEM cab has been removed and replaced with a virtual presentation projected on the screen. Auditory feedback is provided through the OEM four-speaker stereo system. Sound intensity varies with acceleration, braking, and movements off the road. Finally, kinetic feedback is provided through the steering wheel and vibrations from the auditory system. The steering wheel is equipped with a force feedback system that provides increasing levels of resistance as the participant turns the wheel away from the midline. The vehicle's acceleration and braking performance is calibrated in a manner consistent with Volkswagen specifications for the New Beetle (*Motor Trend*, 1999).

The simulator records parameters that measure a wide array of driver and vehicular performance. Computed variables include simulation time, distance traveled, velocity (longitudinal and lateral), acceleration (overall, braking, throttle, and lateral), lane position, roadway curvature, vehicle heading, current transmission gear, and steering wheel angle. Raw measures include encoder counts for the brake, the accelerator, the steering wheel, the directional signal (right or left), the traffic light signal state, and the horn indicator. In addition to the raw and computed measures, the simulator records a running tally of different types of tickets and crashes. Finally, information about other vehicles in the simulation system (e.g., number of cars, range, speed, and lane position) is recorded as a single multidimensional parameter.

Although the simulator clearly resides in a laboratory setting, it provides experimental participants with many elements of a realistic driving environment. The participants sat in an actual vehicle, where the layout and location of controls accord with their mental models of what is usual for automobiles. The participants' responses during the simulation were based on the same kinds of movements and reactions people have when they drive. A steering wheel, an accelerator, and a brake controlled the car's virtual movements, just as they would in a real vehicle. The participants were not required to use unfamiliar equipment or new or different physical movements to direct the vehicle (as a joystick or mouse-controlled simulation might require). The simulator is also equipped to provide drivers with physical feedback to mimic what they would experience were they actually driving, although there is no movement associated with the simulator itself.

The simulation itself requires participants to drive on a road and to deal with obstacles drivers normally encounter, such as other traffic and pedestrians. The task at hand is thus a familiar one—to keep a car moving along a road—not an unfamiliar one, such as a video game scenario might be to some participants. The view projected onto the screen provides drivers with a relatively high resolution environment, and the rearview mirror display accurately depicts what drivers have seen. The range of measurements recorded by the simulator provides a detailed and nuanced picture of drivers' actions and reactions in response to the scenario. Because the physical setting for the experiment and the nature of the simulation itself so closely resemble what drivers in a real environment might confront, there is high face validity associated with the measures. In short, we argue that, for example, participants' braking time in reaction to seeing a stop sign in the simulation is a good measure of their braking time in response to seeing a stop sign in real driving. As one participant noted in the postsimulation questionnaire, "it was a great simulation, felt almost lifelike."

Convergent and discriminant validity. We provide three analyses to establish aspects of the convergent validity of the driving simulation measures empirically. First, we ask

Table 4
Convergent Validity Analysis of Driving Simulation Measures

	Simulator Measures					
Survey Measures	Traffic Weaving	Speeding	Pause Time at Stop Signs	Lane Variability	Mean Control Velocity	Total Accidents
Drove through traffic switching quickly back and	2.543^{**}					
forth between lanes	(1.054)					
Total speeding		0.316^{***}			7.358^{**}	
tickets		(0.102)			(2.697)	
Slowed but failed			-0.274^{*}			
to stop at a stop sign			(0.145)			
Pass vehicle in no-				-0.141^{**}		
passing zone				(0.064)		
Five year accident						0.407^{**}
history						(0.186)
Constant	98.221^{***}	0.154^{***}	3.692^{***}	1.302^{***}	-3.834^{***}	-0.340
	(4.258)	(0.039)	(0.531)	(0.300)	(1.035)	(0.271)
Number of cases	41	39	41	41	39	41
Adjusted R^2	0.108	0.184	0.061	0.089	0.145	0.077

Note—Table entries in the first five columns are unstandardized ordinary least squares regression coefficients with standard errors in parentheses; entries in the last column are probit coefficients with robust standard errors in parentheses. Empty cells indicate that the variable was not included in the analysis. For the last column, the adjusted R^2 value reported is actually the pseudo-R^2 value. s$^*p < .10$. $^{**}p < .05$. $^{***}p < .01$.

how well a number of different driving simulation measures relate to self-reported driving behaviors from questionnaire data collected prior to the simulation. Second, for evidence of concurrent validity, we compare accident involvement in the driving simulation between participants diagnosed with ADHD and controls. Finally, we construct a multitrait–multimethod matrix in order to investigate discriminant and convergent validity. A multitrait–multimethod matrix enables us to compare how well indicators of the same concept measured using different methods agree and to compare the extent to which measures of different concepts disagree (Campbell & Fiske, 1959; Judd & McClelland, 1998; Kidder & Judd, 1986).

Table 4 suggests that there is convergent validation across the six different measures from the simulation data that we examine here. In all cases, the coefficients from the survey items regressed on the simulation measures are statistically significant and in the predicted direction. For example, the participants who reported having received more speeding tickets were more likely both to violate the speed limit more often during the simulation and to have driven, on average, faster in control periods during the simulation. Those who reported that they had been involved in more accidents in the 5 years prior to the simulation experiment were more likely to be involved in a crash in the simulation. That the self-reported driving behaviors from the survey are significantly associated with people's measured behaviors in the simulation increases our confidence that the measures collected from the simulation are valid indicators of people's real driving behaviors.

Previous research has found that individuals with ADHD are more likely to report being involved in accidents than are controls (Barkley et al., 1993; Weiss, Hechtman, Perlman, Hopkins, & Wener, 1979). To establish concurrent validity, we would expect to find the same pattern in the simulation data. To explore whether simulation behavior was consistent with this research, we ran a probit analysis of diagnosis status on a dichotomous dependent variable representing whether the participant had been involved in one or more accidents over the course of the simulation. The results of this analysis, shown in Table 5, are consistent with results from existing research. The participants with ADHD were more likely to be involved in accidents over the course of the simulation than were controls, providing support for concurrent validity among the simulation measures.

A multitrait–multimethod matrix for the simulator and survey data we use is presented in Table 6. The questions here are the extent to which correlations between indicators of the same concept measured in different ways are greater than zero (thus establishing convergent validity) and the extent to which these correlations are greater than methods effects—greater than the correlations of indicators of different concepts measured with the same method (establishing discriminant validity).

The results in Table 6 suggest that the measures from the simulator possess convergent validity. All six of the correla-

Table 5
Concurrent Validity Analysis of Accidents in the Driving Simulation

	Accidents
Type	$.688^{*}$ (.405)
Constant	$-.303$ (.282)
Number of cases	41
Pseudo-R^2	.053

Note—Table entries are probit coefficients with robust standard errors in parentheses. $^*p < .10$.

Table 6
Multitrait–Multimethod Correlation Matrix of Driving Behavior Measures

		Survey Measures					Simulator Measures					
		Drove in Traffic Switching Between Lanes	Total Speeding Tickets	Slowed But Failed to Stop at a Stop Sign	Pass Vehicle in No-Passing Zone	Five-Year Accident History	Traffic Weaving	Speeding	Mean Control Velocity	Pause Time at Stop Signs	Lane Variability	Total Accidents
Survey Measures	Drove in traffic switching between lanes	1.0										
	Total speeding tickets	*−.487****	1.0									
	Slowed but failed to stop at a stop sign	*.093*	*−.347***	1.0								
	Pass vehicle in no-passing zone	*.091*	*−.343***	*.276**	1.0							
	Five-year accident history	*−.264**	*.199*	*−.105*	*.124*	1.0						
Simulator Measures	Traffic weaving	**.360****	−.496***	.248	−.048	−.113	1.0					
	Speeding	−.064	**.452*****	−.434***	−.206	−.204	***−.518******	1.0				
	Mean control velocity	−.126	**.424*****	−.341**	−.129	−.123	***−.591******	***.872******	1.0			
	Pause time at stop signs	−.068	−.001	**−.290***	.087	−.237	***−.227***	***−.059***	***.022***	1.0		
	Lane variability	.050	.070	−.253	**−.336****	−.193	***−.276****	***.387*****	***.274****	***.058***	1.0	
	Total accidents	−.105	.338**	−.208	−.052	**.317****	***.027***	***.045***	***.047***	***−.079***	***.186***	1.0

Note—Table entries are Pearson's bivariate correlation coefficients. Correlations in boldface represent the same trait measured with different methods. Correlations in italics are different traits measured with survey methods. Correlations in boldface italics are different traits measured with simulation methods. $^{*}p < .10$. $^{**}p < .05$. $^{***}p < .01$.

tions between the same trait measured with different methods are significant ($p < .10$), indicating that these correlations are statistically greater than zero. The criteria for discriminant validity are met to some degree, although the evidence is not as great as that for convergent validity. Each of the convergent validity correlations is greater than a majority of its correlations with the other indicators (as Campbell & Fiske, 1959, describe it, the convergent validity correlations are greater than other values in the associated columns and rows in the heterotrait–heteromethod matrix). Yet 5 of the 10 correlations representing different traits measured with survey methods, and 6 of the 14 correlations representing different traits measured with the simulation, attain conventional levels of statistical significance ($p < .10$). While the average size of the relationship between items measuring the same concept is greater than those measuring different concepts with the same method, the difference ideally would be greater than it is. The average of the absolute values of the correlation coefficients for the same trait measured with different methods is .363, whereas that for different traits measured with survey methods is .233, and that for different traits measured with simulation methods is .200. Overall, the average correlation for different traits measured with different methods is .214. Finally, although the patterns of relationships among the variables are not the same across blocks of the matrix, there is some degree of similarity in terms of patterns of statistical significance and relative correlation sizes. Taken together, the results here support claims of convergent validity. The evidence for discriminant validity for the simulation measures we consider is not as strong as that for convergent validity, although the different criteria to establish discriminant validity are met to some degree.

DISCUSSION

In this study, we report an effort to establish the validity of measures of driving behavior collected during a simulation scenario using self-reported survey indicators of driving behavior. The results we show here suggest that the measures we consider are valid indicators of the behaviors of interest. The experimental setting contains features that increase the face validity of the data collected, and the three analyses of convergent and discriminant validity provide further support for the argument about the validity of the simulation measures. While there were positive correlations associated with measures of different concepts with the same method, these were smaller than correlations between measures of the same concept using different methods. Although the convergent validity analysis in Table 4 and the correlations in Table 6 fall far short of perfection, each of the relationships we examine is in the expected, predicted direction, and the sizes of the effects are such that our claims of validity are supported. Similarly, our results are consistent with previous research on drivers with ADHD, adding support for concurrent validity of the measures.

There are several limitations to the study. We had a relatively small number of participants in the experiment, and some of the questionnaire items do not map neatly onto different measures collected during the simulation. We also establish support for validity empirically with a relatively small number of different indicators from the simulation. We argue, however, that these indicators represent a number of different types of behaviors in the driving simulation, and the fact that they are different from each other supports a broader claim about the ability to conclude that the measures from the simulation are valid measures of real driving behaviors.

Taken together, the results here suggest that self-reported survey items present a viable alternative means to assess the validity of measures collected from simulation scenarios. In combination with simulation data, they present the opportunity to consider both the convergent and the discriminant measurement validity of different indicators.

REFERENCES

Allen, R. W., Park, G., Rosenthal, T. J., & Aponso, B. (July, 2004). *A process for developing scenarios for driving simulations*. Paper presented at the IMAGE 2004 Conference, Scottsdale, AZ.

Alm, H., & Nilsson, L. (1995). The effects of a mobile telephone task on driver behaviour in a car following situation. *Accident Analysis & Prevention*, **27**, 707-715.

Barkley, R. A., Guevremont, D. C., Anastopoulos, A. D., DuPaul, G. J., & Shelton, T. L. (1993). Driving-related risks and outcomes of attention deficit hyperactivity disorder in adolescents and young adults: A 3- to 5-year follow-up survey. *Pediatrics*, **92**, 212-218.

Ben-Yaacov, A., Maltz, M., & Shinar, D. (2002). Effects of an in-vehicle collision avoidance warning system on short- and long-term driving performance. *Human Factors*, **44**, 335-342.

Berg, D., Raugi, G. R., Gladstone, H. B., Berkley, J., Weghorst, S., Ganter, M., & Turkiyyah, G. (2001). Virtual reality simulators for dermatologic surgery: Measuring their validity as a teaching tool. *Dermatologic Surgery*, **27**, 370-374.

Biederman, J., & Faraone, S. V. (2004). The Massachusetts General Hospital studies of gender influences on attention-deficit/hyperactivity disorder in youth and relatives. *Psychiatric Clinics of North America*, **27**, 225-232.

Blana, E., & Golias, J. (2002). Differences between vehicle lateral displacment on the road and in a fixed-base simulator. *Human Factors*, **44**, 303-313.

Boyle, J. M., Dienstfrey, S. J., & Sothoron, A. (1998). *National survey of speeding and other unsafe driving actions* (Rep. No. DOT HS 808 749). U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, DC.

Brown, I. D., Tickner, A. H., & Simmonds, D. C. (1969). Interference between concurrent tasks of driving and telephoning. *Journal of Applied Psychology*, **53**, 419-424.

Brunnauer, A., Laux, G., Geiger, E., & Möller, H. J. (2004). The impact of antipsychotics on psychomotor performance with regards to car driving skills. *Journal of Clinical Psychopharmacology*, **24**, 155-160.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, **56**, 81-105.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Carr, D. B., Duchek, J., & Morris, J. C. (2000). Characteristics of motor vehicle crashes of drivers with dementia of the Alzheimer type. *Journal of the American Geriatric Society*, **48**, 100-102.

Cox, D. J., Humphrey, J. W., Merkel, R. L., Penberthy, J. K., & Kovatchev, B. (2004). Controlled-release methylphenidate improves attention during on-road driving by adolescents with attention-deficit/hyperactivity disorder. *Journal of the American Board of Family Practice*, **17**, 235-239.

Cox, D. J., Merkel, R. L., Penberthy, J. K., Kovatchev, B., & Hankin, C. S. (2004). Impact of methylphenidate delivery profiles on driving performance of adolescents with attention-deficit/hyperactivity disorder: A pilot study. *Journal of the American Academy of Child & Adolescent Psychiatry*, **43**, 269-275.

Cox, D. J., Penberthy, J. K., Zrebic, J., Weinger, K., Aikens, J. E., Frier, B., et al. (2003). Diabetes and driving mishaps. *Diabetes Care*, **26**, 2329-2334.

Dawson, S. L., & Kaufman, J. A. (1998). The imperative for medical simulation. *Proceedings of the IEEE*, **86**, 479-483.

Decina, L. E., Gish, K. W., Staplin, L., & Kirchner, A. H. (1996). *Feasibility of new simulation technology to train novice drivers* (Pub. No. DTNH22-95-C-05104). Washington, DC: National Highway Traffic Safety Administration.

Döring, B. (1990). A simulation study for analysing pilot's rule-based behavior. *Proceedings of the computer aided system design and simulation conference* (p. 1-19). France: Neuilly Sur Seine.

Godley, S. T., Triggs, T. J., & Fildes, B. N. (2002). Driving simulator validation for speed research. *Accident Analysis & Prevention*, **34**, 589-600.

Hahn, R. W., & Tetlock, P. C. (1999, October). *The economics of regulating cellular phones in vehicles* (Working Paper, 99-9). AEI-Brookings Joint Center for Regulatory Studies, Washington, DC.

Harms, L. (1996). Driving performance on a real road and in a driving simulator: Results of a validation study. In A. G. Gale, I. D. Brown, C. M. Haslegrave, & S. P. Taylor (Eds.), *Vision in vehicles V* (pp. 19-26). Amsterdam: Elsevier.

Hawkins, R., MacKrell Gaglione, M., LaDuca, T., Leung, C., Sample, L., Gliva-McConvey, G., et al. (2004). Assessment of patient management skills and clinical skills of practising doctors using computer-based case simulations and standardised patients. *Medical Education*, **38**, 958-968.

Hoffman, J. D., Lee, J. D., Brown, T. L., & McGehee, D. V. (2002). Comparison of driver braking responses in a high-fidelity simulator and on a test track. *Transportation Research Record*, **1803**, 59-65.

Holland, C. A., Handley, S., & Feetam, C. (2004). *Older drivers, illness and medication* (Road Safety Res. Rep. No. 39). Department for Transport: London.

Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology: Vol. 1* (4th ed., pp. 180-232). New York: McGraw-Hill.

Kaptein, N. A., Theeuwes, J., & van der Horst, R. (1996). Driving simulator validity: Some considerations. *Transportation Research Record*, **1550**, 30-36.

Kidder, L. H., & Judd, C. M. (1986). Research methods in social relations. New York: Holt, Rinehart & Winston.

Klee, H., Bauer, C., Radwan, E., & Al-Deek, H. (1999). Preliminary validation of driving simulator based on forward speed. *Transportation Research Record*, **1689**, 33-39.

Lagarde, E., Chastang, J. F., Gueguen, A., Coeuret-Pellicer, M., Chiron, M., & Lafont, S. (2004). Emotional stress and traffic accidents: The impact of separation and divorce. *Epidemiology*, **15**, 762-766.

Landau, F., Laur, M., Hein, C. M., Srinivasan, R., & Jovanis, P. P. (1994). *A simulator evaluation of five in-vehicle navigation aids* (Publication No. UCD-ITS-RR-94-19). Institute of Transportation Studies, University of California, Davis.

Lawton, R., Parker, D., Manstead, A. S. R., & Stradling, S. G. (1997). The role of affect in predicting social behaviors: The case of road traffic violations. *Journal of Applied Social Psychology*, **27**, 1258-1276.

Lee, H. C., Cameron, D., & Lee, A. H. (2003). Assessing the driving performance of older adult drivers: On-road versus simulated driving. *Accident Analysis & Prevention*, **35**, 797-803.

Lee, J. D. (2004). *Simulator fidelity: How low can you go*? Retrieved November 1, 2004, from www.uiowa.edu/neuroerg/Simulator%20Users%20Group/HFES_SimPanel.htm.

McGaw, D., & Watson, G. (1976). *Political and social inquiry*. New York: Wiley.

McGwin, G., Jr., Sims, R. V., Pulley, L., & Roseman, J. M. (2000). Relations among chronic medical conditions, medications, and automobile crashes in the elderly: A population-based case-control study. *American Journal of Epidemiology*, **152**, 424-431.

McKnight, A. J., & McKnight, A. S. (1993). The effect of cellular phone use upon driver attention. *Accident Analysis & Prevention*, **25**, 259-265.

Motor Trend (1999). *1999 import car of the year: Volkswagen New Beetle*. Retrieved September 13, 2004, from www.motortrend.com/oftheyear/car/112_9902_icoy/value.html.

Murphy, K., & Barkley, R. A. (1996). Prevalence of DSM-IV symptoms of attention deficit hyperactivity disorder in adult licensed drivers: Implications for clinical diagnosis. *Journal of Attention Disorders*, **1**, 147-161.

National Transportation Safety Board (2004). *Special investigation highway report: Medical oversight of noncommercial drivers* (NTSB No. SIR-04/01). Washington, DC: Author.

Nunes, L., & Recarte, M. (2002). Cognitive demands of hands-free-phone conversation while driving. *Transportation Research Part F: Traffic Psychology & Behaviour*, **5**, 133-144.

Parker, D., Reason, J. T., Manstead, A. S. R., & Stradling, S. G. (1995). Driving errors, driving violations and accident involvement. *Ergonomics*, **38**, 1036-1048.

Ranney, T. A., Heydinger, G., Watson, G., Salaani, K., Mazzae, E. N., & Grygier, P. (2002). *Investigation of driver reactions to tread separation scenarios in the National Advanced Driving Simulator (NADS)* (Rep. No. DOT HS 809 523). Washington, DC: National Highway Traffic Safety Administration.

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

Rimmö, P. A., & Hakamies-Blomqvist, L. (2002). Older drivers' aberrant behaviour, impaired activity, and health as reasons for self-imposed driving limitations. *Transportation Research Part F: Traffic Psychology & Behaviour*, **5**, 47-62.

Rizzo, M., Reinach, S., McGehee, D., & Dawson, J. (1997). Simulated car crashes and crash predictors in drivers with Alzheimer disease. *Archives of Neurology*, **54**, 545-551.

Rolfe, J., & Hampson, B. P. (2003). Flight simulation—Viability versus liability issues of accuracy, data and validation. *Aeronautical Journal*, **107**, 631-635.

Rosenstone, S. J., Hansen, J. M., & Kinder, D. R. (1986). Measuring change in personal economic well-being. *Public Opinion Quarterly*, **50**, 176-192.

Sarathy, S., & Higman, J. (1994). Development and validation of the OH-58D Kiowa Warrior High Fidelity Flight Simulation Model. *Proceedings of the annual forum of the American Helicopter Society*, **2**, 905-914.

Sodhi, M., Reimer, B., & Llamazares, I. (2002). Glance analysis of driver eye movements to evaluate distraction. *Behavior Research Methods, Instruments, & Computers*, **34**, 529-538.

Stanton, N. A., Young, M. S., Walker, G. H., Turner, H., & Randle, S. (2001). Automating the driver's control tasks. *International Journal of Cognitive Ergonomics*, **5**, 221-236.

Stein, A. C., Allen, R. W., & Parseghian, Z. (1992, July). *The use of low-cost interactive driving simulation to detect impaired drivers*. Paper presented at the IMAGE VI Conference, Scottsdale, AZ.

Stradling, S. G., & Meadows, M. L. (2000). Highway code and aggressive violations in UK drivers. In *Proceedings of the Global Web Conference on Aggressive Driving Issues* (pp. 1-14). Ontario: Ministry of Transportation of Ontario.

Stutts, J. C., & Wilkins, J. W. (2003). On-road driving evaluations: A potential tool for helping older adults drive safely longer. *Journal of Safety Research*, **34**, 431-439.

Törnros, J. (1998). Driving behavior in a real and a simulated road tunnel—a validation study. *Accident Analysis & Prevention*, **30**, 497-503.

Weiss, G., Hechtman, L., Perlman, T., Hopkins, J., & Wener, A. (1979). Hyperactives as young adults: A controlled prospective ten-year follow-up of 75 children. *Archives of General Psychiatry*, **36**, 675-681.

Zuin, D., Ortiz, H., Boromei, D., & Lopez, O. L. (2002). Motor vehicle crashes and abnormal driving behaviours in patients with dementia in Mendoza, Argentina. *European Journal of Neurology*, **9**, 29-34.

(Manuscript received November 15, 2004;
revision accepted for publication March 29, 2005.)