THE GENERALIZABILITY OF CARDIOVASCULAR RESPONSES ACROSS SETTINGS

WILLIAM GERIN,* NICHOLAS CHRISTENFELD,† CARL PIEPER,‡
DENISE A. DeRAFAEL,§ OLINDA SU,|| STEVEN J. STROESSNER,||
JAMES DEICH‡ and THOMAS G. PICKERING*

(Received 15 October 1996; accepted 5 June 1997)

Abstract—The generalizability of cardiovascular reactivity change scores remains largely unsupported. In previous studies, several factors differed between laboratory and field, making poor lab-to-life correlations difficult to interpret. The present study varied only one parameter between the lab and field: setting. In this study, 24 females were studied on four occasions: twice in the lab (to provide test–retest reliability); once in a classroom; and once at home. After a baseline, subjects performed a math task, while blood pressure and heart rate were monitored. Procedures were identical in all sessions. Blood pressure changes were fairly reliable between the two lab sessions, with r values 0.68 (systolic) and 0.62 (diastolic pressure); however, lab/nonlab correlations were lower (0.47 for SBP; 0.38 for DBP). This suggests that even a minor variation in procedure, such as a change in setting, can affect generalizability; other lab–field differences may have an even greater impact. © 1998 Elsevier Science Inc.

Keywords: Cardiovascular reactivity; Generalizability; Blood pressure; Heart rate.

INTRODUCTION

In a critique of the cardiovascular reactivity paradigm, Pickering and Gerin [1] concluded that, if the role of reactivity in the development of cardiovascular disease was to be adequately evaluated, several aspects of reactivity required more support than was currently available. Since that time, additional evidence in some areas has emerged. For example, a recent report indicates that substantial improvements in the test–retest reliability of cardiovascular change scores can be made if measures are aggregated across various tasks [2]. Little support has been forthcoming, however, concerning the generalizability of laboratory reactivity change scores to the natural environment. Such support is crucial, because the reactivity hypothesis rests on the assumption that reactivity is a stable individual difference, and the rank ordering of individuals, in terms of their reactivity scores, must therefore be preserved

* Cornell University Medical College–The New York Hospital, New York, New York, USA.
† University of California, San Diego, San Diego, California, USA.
‡ Duke University Medical Center, Durham, North Carolina, USA.
§ City College/City University of New York, New York, New York, USA.
|| St. John’s University, New York, New York, USA.
¶ Barnard College, Columbia University, New York, New York, USA.

Address correspondence to: William Gerin, Ph.D., Hypertension Center/Starr Pavilion ST405, Cornell University Medical College/New York Hospital, 525 East 68th Street, New York, NY 10021, USA. Tel: (212) 736-0791; Fax: (212) 279-6281; E-mail: wgerin@med.cornell.edu
across situations. If it is not, then this will limit the usefulness of cardiovascular reactivity change scores as predictors of long-term disease.

Most studies in which generalizability has been examined have had subjects undergo reactivity testing in the laboratory, and then wear an ambulatory monitor, which measures blood pressure and heart rate intermittently, during the subjects' normal activities. The laboratory change scores are then correlated with some measure of variability of the measurements taken in the field, such as the standard deviation. These studies have found, at best, small associations [6–13].

Of course, many elements varied between the laboratory and field situations, and one cannot tell from the small lab–life correlations which of these may have affected reactivity in the field, nor can one tell the extent to which the small lab–life associations are due to less than perfect reliability of measurements in both settings. Only one study has systematically varied a limited number of elements, and examined the association between reactivity measured during two laboratory sessions. Smith and O'Keefe [14] examined the cross-situational consistency of responses by simultaneously varying four elements: setting, experimenter, measurement apparatus, and task. They found significant, but rather small, associations between the sessions. Controlling for baseline, the correlations for systolic and diastolic blood pressure changes were 0.39 and 0.17, respectively. The correlation for heart rate change was 0.43. As would be expected, these correlations are somewhat higher than those found between laboratory change scores and ambulatory standard deviations. However, even in the Smith and O'Keefe study, most of the variance across sessions remains unexplained; and, as with the laboratory–field associations, it cannot be determined from these data which of the varied elements were responsible for attenuating the associations.

There are many possible causes for poor generalizability, as a variety of elements vary between the laboratory and the subjects' natural environment. For example, the tasks that subjects perform in the lab are not the same ones in which they engage in the normal course of their lives. While the researcher intends the laboratory tasks to be conceptually representative of the array of stressors that people really face, there is no evidence at this point to support this. Even simple factors such as postural differences can add error variance, diminishing the lab–life relationship [15].

One factor that necessarily varies between the lab and the natural environment is setting. It may be that responses in the laboratory, taken using the same equipment, with the same subject posture, and using the same stressors, may not be representative of how that individual would respond at home, or at work. The laboratory is a distinctive environment, often filled with medical paraphernalia, and is probably unlike the places in which most subjects spend their days. This study explores whether just changing the setting, while holding all other factors constant, reduces generalizability of responses measured in the laboratory.

There is considerable similarity between the issues underlying the generalizability of blood pressure changes from the laboratory to the natural environment and

---

1 Some positive results have been reported, in studies using different methods; for example, larger associations for some tasks, using intraarterial recording, have been reported [3, 4]. Using a different strategy, Matthews et al. [5] have shown that cardiovascular responses to a laboratory challenge predicted responses to a real-world stressor (giving a speech) among tenth-grade students.
the issue of cross-situational consistency that has received attention in the personality literature. For example, Mischel [16] has pointed out that most studies have shown that behavior measured under one set of conditions was a poor predictor of those behaviors under different conditions, usually showing no greater than a cross-situational correlation of 0.30. For the most part, this literature has not been applied to physiological measurements taken in the laboratory; however, the consistent lack of cross-situational consistency observed suggests that the same problem may exist for physiological stress responses as for other types of more observable behavior.

Classical test theory suggests that when measurements are taken under identical conditions, on two separate occasions, any difference that occurs between those measurements (barring developmental changes) must be due to random error. This is referred to as “test–retest reliability,” and is usually measured using a Pearson correlation. It is assumed that the test–retest condition reflects only random error and that the correlation can, in theory, only weaken when one or more elements of the situation change from one test occasion to the next. When such changes occur, the issue becomes one of “generalizability” (also usually measured using a correlation). Thus, generalizability refers to the extent that a measurement taken under one set of circumstances will be reproducible under one or more different circumstances. The generalizability correlation reflects residual variance that is due to two sources: random error plus the variability due to the change, or changes, from one set of measurements to the next.

The present study was designed to examine the generalizability of cardiovascular responses from the laboratory to nonlaboratory settings. Subjects’ reactivity to mental stress was measured on four occasions: twice in the laboratory; once in a classroom; and once in the subject’s home. Apart from setting, all other factors, including apparatus, task, and experimenter, were maintained constant. This design allows us to evaluate the relationship between reactivity scores when measured under identical conditions (i.e., the test–retest reliability), as well across settings (i.e., generalizability).

METHOD

Overview

Each subject participated in four sessions. These were held on a Monday, Tuesday, Thursday, and Friday, of the same week. Two of the sessions were held in the cardiovascular reactivity laboratory; one session was held in a classroom on campus; and one session was held in the subject’s apartment. Order of settings was counterbalanced across subjects. Each session comprised an initial baseline and a mental arithmetic stressor. During both phases, blood pressure and heart rate were monitored.

Subjects

Twenty-four female students attending a small eastern university participated. The subjects were all between the ages of 17 and 26 years, and all were normotensive (resting blood pressure less than 140/90 mmHg). Subjects participated in exchange for a cash payment, and were asked to refrain from caffeine and nicotine use for at least 2 hours prior to the session. All subjects lived within two blocks of the laboratory.

Recording of blood pressure and heart rate

Blood pressure and heart rate were collected using an A&D Series VII, Model 2421 ambulatory blood pressure monitor (ABPM). This monitor uses both auscultatory and oscillometric methods of blood pressure measurement, although only auscultatory measurements were used in the present study. All measurements can be stored, and later downloaded, to a computer for analysis. The A&D 2421 monitor has
been extensively validated [17–19], and has been found to satisfy both the AAMI and BHS criteria [20]. Because the monitoring was done under the supervision of an experimenter, the occasional “bad reading” was immediately replaced with a subsequent measurement. (A bad reading occurred when the monitor registered an error, either due to the subject’s excessive movement (this occurred on one reading for one subject in the initial baseline phase); or due to poor cuff placement (this occurred on six readings for four subjects, all during the early part of the initial baseline phase; when this occurred, the microphone placement was immediately corrected, and a subsequent measurement was then taken to replace the bad reading).

Task

A 3-minute serial-subtraction task was used with subjects counting backwards, aloud, by 13s. A different starting point was used in each of the four sessions. Subjects were asked to count as quickly and accurately as possible; and were told that the experimenter would correct them if necessary. During the task, the experimenter goaded the subject at periodic intervals (approximately every 20–30 seconds), using the phrases: “try to go a little faster” and “your time is running out.” The timing of these statements was maintained constant across sessions and locations.

This task was selected for two reasons: (1) it is the most widely used mental task in reactivity studies [1]; and (2) it represents an active-coping challenge, which may be representative of a broad class of stimuli with which individuals may come into contact in the real world [21–23].

Counterbalancing

Subjects were randomly assigned to one of six possible orders with four subjects in each. Although there were a total of four sessions, the two laboratory sessions were always contiguous in order to provide estimates of test–retest reliability.

Procedures

Subjects always met the experimenter at the cardiovascular laboratory by appointment. It had been explained previously, however, that the session would take place at a variety of settings, including in the subject’s apartment. Upon arrival at the laboratory, subjects were escorted to one of the settings: inside the laboratory; a classroom located on a different floor of the same building; or to the subject’s apartment (located within two blocks of the laboratory). Subjects were studied at approximately the same time of day in all sessions. Upon reaching the setting, subjects were seated and instrumented using the ABPM arm cuff. Subjects sat upright in a comfortable position, with hands on a table in front of them. Subjects then sat quietly for a 12-minute baseline period. Blood pressure and heart rate measurements were taken at 0, 3, 6, 9, and 12 minutes during this period. At the end of the task, subjects were asked to rate the stressfulness of the task on a Likert-type scale, where 1 = “not at all stressful” and 7 = “extremely stressful.”

Next, the experimenter explained the math stressor (session 1), and the subject performed the task for 3 minutes. During this period, blood pressure and heart rate measurements were taken at 1 and 3 minutes; subjects continued the task during arm-cuff inflation. Two experimenters conducted the study, each subject was always studied by the same experimenter, and experimenters were randomly assigned to location order.

After the final measurement, the arm-cuff was removed. Arrangements were made for the next appointment (sessions 1, 2, and 3) or the subject was paid (session 4).

Manipulation of testing site

Testing site comprised the independent variable. Sessions were held in one of three settings: (1) The cardiovascular laboratory, which contained several obvious “laboratory” items (automated blood pressure monitors, computers, etc.). The room was approximately 10’ × 12’; and well-lit. A one-way mirror was positioned prominently on one wall. The experimenter and subject were seated at a large rectangular table, in the center of the room. (2) A classroom, located on a different floor of the same building as the laboratory. The classroom was considerably larger than the laboratory, had several windows, and contained student desks and a blackboard. Testing was conducted at one of the student desks. (3) The living room of the subject’s apartment; these, of course, varied.

Data reduction and statistical analyses

Five measurements were taken during the baseline phase; the first two of these were discarded to allow for adaptation, and the baseline measure was the mean of the final three measurements. Two measurements were taken during the task phase, and the task measure was represented by the mean of both measurements. The change score was computed as the task level minus the baseline level.
Generalizability of cardiovascular responses

Table I.—Mean (SD) blood pressure and heart rate resting levels and reactivity change scores by setting (N = 24)

<table>
<thead>
<tr>
<th>Setting</th>
<th>Systolic pressure</th>
<th>Diastolic pressure</th>
<th>Heart rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>Change</td>
<td>Baseline</td>
</tr>
<tr>
<td>Lab 1</td>
<td>97.1 (11.2)</td>
<td>12.1 (9.7)</td>
<td>73.0 (4.3)</td>
</tr>
<tr>
<td>Lab 2</td>
<td>94.7 (10.0)</td>
<td>13.7 (8.9)</td>
<td>72.9 (4.8)</td>
</tr>
<tr>
<td>Classroom</td>
<td>97.6 (10.1)</td>
<td>14.5 (7.7)</td>
<td>72.7 (3.7)</td>
</tr>
<tr>
<td>Home</td>
<td>95.3 (11.9)</td>
<td>14.6 (11.0)</td>
<td>72.1 (4.9)</td>
</tr>
</tbody>
</table>

Analysis of variance (ANOVA) procedures were performed on the cardiovascular measures, with phase (baseline, task) and location serving as a within-subject factors.

Pearson correlations were computed based on the cardiovascular means and change scores taken at each of the locations. These were provided to represent the test–retest reliability, as well as the generalizability across different settings.

RESULTS

Effects of setting on blood pressure and heart rate resting levels and changes

Table I shows the mean cardiovascular resting levels and change scores for each setting. As the table shows, the mean resting levels and change scores did not vary a great deal across settings compared to the standard deviations of the measurements within each setting. Thus, for example, the range among systolic blood pressure levels was less than 3 mmHg; the lowest measurement was 94.7 mmHg (lab 2) and the highest was 97.6 mmHg (classroom), compared to within-setting standard deviations averaging 10.8 mmHg. Repeated-measures ANOVAs were used to estimate the effects of the task and of location on blood pressure and heart rate. The ANOVA indicated that, as expected, the task had a significant effect on all three outcomes, with Fs(1, 23) = 84.62, 80.23, and 112.39, for systolic blood pressure, diastolic pressure, and heart rate, respectively, all ps<0.0005. However, neither the main effects of location, nor the interaction between location and phase, were significant [highest F(3, 69) = 1.71, for the main effect of location on systolic pressure].

Test-retest reliability of cardiovascular levels and change scores

Pearson test–retest correlations are shown in Table II (baseline levels) and Table III (change scores). The correlations between each of the lab sessions and the classroom and the home settings are also provided. Table II shows that, for levels, systolic blood pressure was quite reliable, with a test–retest correlation of 0.81. Test–retest correlations were somewhat lower for diastolic pressure (r=0.63) and heart rate (r=0.68). Table II also shows that correlations between the lab settings and the classroom and home settings were similar to the test–retest correlations, indicating good generalizability of levels.

For change scores, the pattern is different. Table III shows the Pearson correlations between the change scores. For the two blood pressure measures, the test–retest correlations are fairly strong, with rs=0.68 and 0.62, for systolic and diastolic pressures, respectively. For heart rate, however, the test–retest correlation between the change scores is almost zero.
Table II.—Pearson correlations of blood pressure and heart rate baselines among the four locations \((N = 24)\)

<table>
<thead>
<tr>
<th></th>
<th>Systolic blood pressure</th>
<th>Diastolic blood pressure</th>
<th>Heart rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lab 1</td>
<td>Lab 2</td>
<td>Classroom</td>
</tr>
<tr>
<td>Lab 2</td>
<td>0.81a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classroom</td>
<td>0.85a</td>
<td>0.89a</td>
<td>0.78a</td>
</tr>
<tr>
<td>Home</td>
<td>0.87a</td>
<td>0.77a</td>
<td>0.78a</td>
</tr>
</tbody>
</table>

\(^a\ p < 0.01\) (two-tailed).

Table III.—Pearson correlations of blood pressure and heart rate changes among the four locations \((N = 24)\)

<table>
<thead>
<tr>
<th></th>
<th>Systolic blood pressure</th>
<th>Diastolic blood pressure</th>
<th>Heart rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lab 1</td>
<td>Lab 2</td>
<td>Classroom</td>
</tr>
<tr>
<td>Lab 2</td>
<td>0.68b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classroom</td>
<td>0.47a</td>
<td>0.55b</td>
<td></td>
</tr>
<tr>
<td>Home</td>
<td>0.44a</td>
<td>0.32</td>
<td>0.45a</td>
</tr>
</tbody>
</table>

\(^a\ p < 0.05\) (two-tailed); \(^b\ p < 0.01\) (two-tailed).
Table IV.—Average Pearson correlations corrected for attenuation due to unreliability for systolic pressure (SBP), diastolic pressure (DBP) and heart rate (HR) levels and changes ($N = 24$)

<table>
<thead>
<tr>
<th>Average Person $r$</th>
<th>Baseline measurements</th>
<th>Change scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SBP</td>
<td>DBP</td>
</tr>
<tr>
<td>Uncorrected</td>
<td>0.84</td>
<td>0.60</td>
</tr>
<tr>
<td>Corrected</td>
<td>1.00$^a$</td>
<td>0.95</td>
</tr>
</tbody>
</table>

$^a$ Because the average correlation is greater than the reliability, the corrected coefficient is $>1$, and has been set to 1.00.

$^b$ The reliability of this measure was extremely poor, thus the corrected coefficients will be highly unstable.

For the blood pressure change scores, the correlations between the laboratory and the nonlab sessions are, for the most part, substantially smaller than the test–retest correlations. These data suggest that blood pressure changes taken in the laboratory may not generalize very well across different settings.$^2$

**Attenuation of generalizability due to unreliability**

To estimate the theoretical magnitude of the lab–nonlab correlations had the measurements been perfectly reliable, the Spearman correction for attenuation was applied. This correction makes use of the reliability of the measurement in question to estimate the extent to which an association between that measurement and some other will be smaller than it would had the reliability been perfect (i.e., 1.0). The resultant association is then "corrected" by inflating the coefficient so that it provides an estimate of the association had the measurement in question been perfectly reliable. The test–retest correlations served as the estimates of reliability required to solve these equations. Table IV shows the uncorrected and corrected correlations. To simplify, the averages of the correlations among the various locations, excluding those between the two lab sessions, are shown (converted using the Fisher $Z$). Note that, due to the poor reliability of heart rate changes, corrected scores for this parameter are not shown. Table IV indicates that once the unreliability is accounted for, generalizability for the blood pressure changes attains moderate levels, with correlations of 0.66 (systolic blood pressure) and 0.56 (diastolic pressure).

**Effects of test site on performance and self-reported stress**

We examined the effects of the different testing sites on measures that might be related to blood pressure and heart rate changes, specifically performance (rates of

$^2$ Order was not completely counterbalanced, because the two lab sessions were always contiguous. Thus, a potential confound exists since the test–retest correlations were always one day apart, and this is not always true of the generalizability (i.e., lab/nonlab) correlations. To determine if number of days between sessions influences the correlations, we have compared generalizability coefficients for the change scores, collapsed across blood pressure and heart rate measures, as a function of number of days apart. The correlations were: 0.32 (1 day); 0.31 (2 days); and 0.31 (3 days). For the two blood pressure measures only, the averaged correlations were: 0.43 (1 day); 0.41 (2 days); and 0.43 (3 days). Thus, it appears that the number of days separating sessions did not systematically affect the associations.
Table V.—Mean (sd) subtraction rates, errors and self-reported stress by setting (N = 24)

<table>
<thead>
<tr>
<th>Setting</th>
<th>Subtraction errors</th>
<th>Number of problems solved</th>
<th>Self-reported stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab 1</td>
<td>3.4 (1.2)</td>
<td>69.0 (7.4)</td>
<td>4.1 (1.1)</td>
</tr>
<tr>
<td>Lab 2</td>
<td>4.1 (1.2)</td>
<td>71.6 (6.9)</td>
<td>3.6 (1.4)</td>
</tr>
<tr>
<td>Classroom</td>
<td>3.1 (1.5)</td>
<td>71.4 (7.1)</td>
<td>3.7 (1.2)</td>
</tr>
<tr>
<td>Home</td>
<td>3.8 (1.4)</td>
<td>70.1 (6.4)</td>
<td>3.4 (1.2)</td>
</tr>
</tbody>
</table>

subtraction and number of errors) and self-reports of stress. Table V, which shows the means for each of these measures, at each location, indicates that there was no systematic effect of the test site on any of the measures, and this is substantiated by the nonsignificant ANOVAs computed for each measure.

DISCUSSION

Baseline levels

The test–retest correlation was quite strong for systolic blood pressure baseline levels, and somewhat weaker for diastolic pressure and for heart rate. When examined across all four settings, the generalizability for levels was similar to the test–retest reliability for each parameter, indicating that a measure taken in the laboratory (averaged, in this case, across three measurements) was a fairly good representation of that measure taken in other settings, under similar circumstances (e.g., by the same trained researcher, using the same equipment, the same task, and so on). The generalizability for diastolic pressure and heart rate levels was somewhat poorer than for systolic pressure, but was still fairly high.

Change scores

For blood pressure, the test–retest correlations were in the moderate range; the correlations between the lab and nonlaboratory settings were smaller. For heart rate, the test–retest correlation was very poor. Taken together, the correlations suggest that a change in setting, even with all other aspects of the test situation held constant, attenuates generalizability.

After correcting for unreliability, the correlations between measurements taken in the laboratory and those taken outside the lab become somewhat stronger. Even the corrected correlations, however, leave most of the variance in nonlaboratory change scores unexplained by the laboratory measurements. It seems clear that the poor generalizability of the change scores is not simply due to unreliability of the measures. In practice, the relationships will, of course, be even weaker. It appears that the simple change of setting is capable of producing changes in reactivity, with different individuals showing greater reactivity in different settings.

As noted earlier, a parallel exists between the present results and the findings that led to the debate over the consistency of personality in general [16]. Repeatedly it has been found that, even with personality measures that provide high test–retest reliability, apparently minor changes in the situation can substantially attenuate the correlation. For example, in their classic study, Hartshorne and May [24] found that how much students cheat in the classroom was stable from day to day. However, it
was a very poor predictor of how much they cheated at home, or cheated in other settings. The prevalence of the person-by-situation interaction in a variety of domains weakens the ability to make predictions based on personality. Our results suggest that the reactivity hypothesis, which depends on cross-situational consistency in the same way as any other personality measure, must also confront this challenge.

Conclusions

One desirable characteristic of reactivity testing is that it can be done easily and under controlled conditions. However, there is a trade-off between convenience and the extent to which values measured in the laboratory adequately represent the person’s response in the natural environment. The results of the present study suggest that reactivity testing conducted in the laboratory provides only a moderately accurate representation of the individual’s “true” reactivity (to that particular stressor, measured using that particular apparatus, etc.).

It is our guess that the specific challenges and stressors actually encountered represent the most significant changes between the lab and the natural environment. Laboratory stressors are selected, among other reasons, to represent a broad range of real-life challenges. In the present study, for example, one reason for the use of the mental arithmetic task was its active coping nature and, presumably, active coping to overcome stress is a fairly common experience for most individuals. How well the laboratory stressor actually represents this broad range of challenges, however, remains to be demonstrated. Thus, the results of the present study suggest that the blood pressure response observed in the laboratory is at best only moderately generalizable to nonlaboratory situations.

To learn to what precise domains we can legitimately generalize, factors which may be sources of variability must be investigated. This is crucial information in the design of future studies, because those factors that do not much matter (i.e., account for a significant proportion of the variance in the outcome measurements) may safely be ignored; and those factors that do produce variance may either be controlled, or at least measured. One example of this concerns the time of day that the testing occurs. We have found in our own laboratory that the stability of blood pressure responses is affected little by different times of day (i.e., morning vs. afternoon). However, other measures, such as cortisol, are heavily influenced by the time of day during which testing occurs [25]. Thus, it may be said that, for blood pressure reactivity, the response generalizes across time of day, but for cortisol response, generalizability across time of day is poor. Knowledge of these associations, and lack of associations, are crucial to the design of reactivity studies.

The present study demonstrates that if we are to find predictive power from the laboratory to the natural environment, there is no dimension of variability so trivial that it can be dismissed without investigation. If simply changing the location of the test site can reduce the lab-to-life associations, then altering more significant aspects of the test situation, such as the task or the subject’s motivation, is likely to do even more damage to the stability of reactivity as an individual difference.

Acknowledgments—Funding for this study was provided by National Institutes of Health, Grant HL.48240.
REFERENCES


