The application of generalizability theory to blood pressure resting levels and mental stress responses
William Gerin, Nicholas Christenfeld* and Carl Pieper†

Background Blood pressure measurements taken in the clinic or laboratory are assumed to generalize to the world outside. This is true both of casual blood pressure measurements and of changes in blood pressure responses to stress. Such generalizability is crucial to the usefulness of blood pressure measurements as predictors of long-term disease. In previous generalizability studies, several factors differed between clinic/laboratory and field, making it difficult to interpret the poor laboratory-life associations. The present study varied only one parameter between the laboratory and the field setting.

Subjects and methods Twenty-four women were studied on four occasions: twice in the laboratory, once in a classroom, and once at home. After a resting baseline, the subjects performed a mathematics task, while blood pressure and heart rate were monitored. Only the setting was varied across sessions.

Results Test–retest correlations were 0.81 for systolic blood pressure levels (SBP) and 0.61 for diastolic blood pressure levels (DBP). Generalizability (G) coefficients for blood pressure levels were approximately the same as the reliabilities (0.82, SBP; 0.59, DBP), indicating that the change in location did not affect resting levels. However, for change scores, the reliabilities were higher than the G coefficients. Test–retest correlations were moderate: 0.68 (SBP) and 0.62 (DBP); G coefficients were 0.47 (SBP) and 0.36 (DBP), indicating that the generalizability of change scores suffered due to the change in test location.

Conclusions A minor variation in procedure, such as a change in setting, has little effect on the generalizability of blood pressure resting levels, but a substantial effect on stress-response changes. Other laboratory-field differences may have an even greater impact on generalizability.

Introduction Blood pressure measurements taken in the clinic have long been regarded as the standard for the prediction of future hypertension [1]. The usefulness of any predictor, however, depends to some extent on the stability of the measurements both in terms of the test–retest reliability of the measurements (taken on more than one occasion under identical conditions) and the ecological validity (taken under a variety of conditions). The latter condition is illustrated by the problem of ‘white-coat hypertension’ [2], in which patients’ blood pressures are elevated when taken in the clinic, but not when measured outside the clinic, as at the patient’s workplace. In this situation, the measurements taken in the clinic poorly represent the individual’s blood pressure level for prognostic purposes. In addition to casual blood pressure measurements taken in the clinic, blood pressure responses to laboratory stressors (cardiovascular reactivity) have also been proposed as a means of predicting future blood pressure levels [3]. In a critique of the cardiovascular reactivity paradigm, Pickering and Gerin [4] concluded that, if the role of reactivity in the development of cardiovascular disease were to be evaluated adequately, it would be necessary to demonstrate the stability of such responses in the natural environment, as well as in the laboratory. Little support for generalizability of these measures, however, has been forthcoming. 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 across situations. If it is not, then this will limit the useful-
ness of cardiovascular reactivity change scores as predictors of long-term disease.

Several studies have focussed on the generalizability of blood pressure resting levels (similar to casual blood pressures taken in the clinic) and of blood pressure stress responses (change scores). For the most part in these studies, one or more resting blood pressures are taken in the laboratory, and the subjects are then often exposed to a stressor of some sort; additional blood pressure measurements are taken during the stressor period, and a change score is computed by subtracting the resting level (the mean of the resting measurements) from the task level (the mean of the task measurements). The subject then wears an ambulatory monitor, which measures blood pressure and heart rate intermittently, during the subject's normal activities. The laboratory resting levels are then correlated with the mean blood pressures taken in the field; and the laboratory change scores are then correlated with some measure of variability of the measurements taken in the field, such as the SD. These studies have found, overall, fairly high correlations for levels taken in the laboratory and the field, but small associations for change scores taken in the laboratory with measures of field variability [5-12], although some positive results have been reported using different methods (for example, larger associations for some tasks, using intraesophageal recording, have been reported [13,14]; using a different stressor, Matthews et al. [15] have shown that cardiovascular responses to a laboratory challenge predicted responses to a real-world stressor (giving a speech) among tenth-grade students.

Of course, many elements vary between laboratory and field situations, and one cannot tell from the small laboratory-life correlations which of these may have affected reactivity in the field; nor can one tell the extent to which the small laboratory-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 reactivities measured during two laboratory sessions. Smith and O'Keefe [16] examined the cross-situational consistency of responses by simultaneously varying four elements: setting, experimentor, measurement apparatus, and task. They found significant, but rather small, associations between the sessions. Controlling for baseline, the authors found correlations for systolic and diastolic blood pressure changes of 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 SD. However, even in Smith and O'Keefe's study [16], most of the variance across sessions remains unexplained and, as with the laboratory-field associations, these data cannot show which of the varied elements was responsible for attenuating the associations.

There are many possible causes for poor generalizability of laboratory change scores, as a variety of elements vary between the laboratory and the subjects' natural environment. For example, the tasks that subjects perform in the laboratory 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 supporting evidence for this. Even simple factors such as postural differences can add error variance, diminishing the laboratory-life relationship [17].

One factor that necessarily varies between the clinic or laboratory and the natural environment is setting. It may be that responses in the laboratory, taken using the same equipment, with the same subject posture, may not be representative of the way that individual responds at home, or at work. The clinic and laboratory are distinctive environments, often filled with medical paraphernalia, and are probably unlike the places in which most subjects spend their days. In this study we explored the question of whether just changing the setting, while holding all other factors constant, can reduce generalizability of levels and change scores measured in the laboratory. In practice, of course, many elements differ between the clinic or laboratory and the natural environment; thus, the present study represents a "best case" scenario, in which only a single element is known to vary between laboratory and field.

Classical test theory suggests that when measurements are taken under identical conditions on two separate occasions any difference that occurs between those measurements (baring developmental changes) must be due to random error. This is referred to as test-retest reliability. 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 arising from 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 levels and responses from the laboratory to non-laboratory settings. Subjects' resting levels and reactivities to mental stress were 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, stressor, and experimenter, were maintained constant. This design allowed
us to evaluate the relationship between testing levels, and between reactivity scores, when measured under identical conditions (test–retest reliability) and across settings (generalizability).

The reliability of measurements is usually assessed using an interclass correlation, such as the Pearson r, or using differences between pairs of measurements. However, generalizability theory, or G theory [18], provides an alternative set of methods, which are used to estimate the degree to which a particular score is representative of the individual's 'true' score, across a wide variety of situations.

For purposes of assessing generalizability, G coefficients hold several advantages over the more often-used Pearson correlations. One advantage is that the analysis of variance model on which the G coefficient is based can simultaneously evaluate more than one potential source of variability; the Pearson correlation can examine only one factor at a time. During ambulatory blood pressure monitoring, for example, a researcher may want to evaluate the effects of changes both in posture and in setting on blood pressure measurements. A Pearson correlation between the measurements taken in the laboratory and those taken in the field may indicate that one or both of these factors did make a difference (i.e., the correlation is small, indicating poor generalizability). However, which of those factors is responsible for the poor generalizability cannot be determined from the Pearson correlation. Using a G theory model, however, is a way of estimating the separate effects of each factor as well as their interaction.

Not only does G theory evaluate more than one factor at a time, a second advantage is that more than two levels of a factor may be simultaneously evaluated using G theory models. The Pearson correlation can only examine pairs of levels. Therefore, in the present study, we were able to examine the effect of a change in setting across three levels of this factor: laboratory, classroom, and home.

Finally, because G theory coefficients are based on intraclass, rather than interclass correlations, they are sensitive to differences in the mean baseline and change scores, as a function of different locations, and will tend to be smaller when such differences represent a substantial source of variability. However, Pearson correlations will not be weakened when this is not the case: the Pearson r only provides information about the relative ordering of individuals across the two settings.

G coefficients are comparable to the reliability coefficients used in classical test theory. The computational procedures are derived from generalizability theory [18–20]. G theory extends the intraclass correlation to allow for more than one source of variance, yielding a G coefficient. In particular, the analyses described here represent a generalizability study as opposed to a decision study [18,19].

An excellent example of a decision study as applied to blood pressure measurements is already available [19].

The G coefficients are derived from the analysis of variance source table. In these analyses, persons and settings are considered to be random effects for the purpose of providing variance estimates. In a random-effects analysis, the issue is not whether a particular level of the independent variable is significantly different from another level, but instead revolves around how much of the total variance in the measurements is due to the different levels of the independent variable (in the present case, to settings).

The analysis of variance source table provides the variances used in the computation of the estimated variance components for each facet of the study (in generalizability theory terms, a ‘factor’ is analogous to a ‘factor’ in a fixed-effects analysis of variance study; the levels of the factor in the present study are the various settings). In the present design, there are four potential sources of variance: variance due to persons (this is regarded as the ‘true’ variance, because it represents the extent to which the measurements are useful for differentiating among individuals); variance due to the main effect of setting; variance due to the person-by-settings interaction; and variance due to error. The last two sources, however, are confounded and therefore indistinguishable in the analysis of variance table. With only one observation in each cell of the table, we do not know, after accounting for the first two sources, whether differences between settings reflect the person-by-settings interaction, the random or unidentified sources of variability, or both. Consequently, they are lumped together as the residual.

It is particularly important to understand the practical implications of two of the effects. First is the main effect of settings. If the effect of settings accounts for a large portion of the total variance, this will indicate that different settings have systematically different effects on blood pressure and heart rate. For example, on average, blood pressure levels might be higher in the classroom than in the home setting. Second is the person-by-settings interaction. If this term accounts for a large proportion of the variance, it indicates that some individuals are more reactive in some settings than in others, and that the settings which produce greater reactivity are different for different individuals. If the person-by-settings interaction accounts for a large proportion of the variance, this suggests that cardiovascular reactivity as an individual difference measure will generalize poorly across settings.

Subjects and methods
Overview
Each subject participated in four sessions. These were held on Monday, Tuesday, Thursday, and Friday of the same week. Two of the sessions were held in the cardio-
vascular reactivity laboratory, one session was held in a classroom on campus, and one session was held in the subject's apartment. The order of settings was counterbalanced across subjects. Each session comprised an initial resting baseline and a mental arithmetic stressor. During each phase, both blood pressure and heart rate were monitored.

Subjects
Twenty-four female students attending a small eastern university participated. The subjects were all aged 17–26 years, and all were normotensive (resting systolic/diastolic blood pressure less than 140/90 mmHg). They participated in exchange for a cash payment, and were asked to refrain from caffeine and nicotine use for at least 2 h before each 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 (Tokyo, Japan) ambulatory blood pressure monitor. This monitor uses both auscultatory and oscillometric methods of blood pressure measurement. All measurements can be stored, and later downloaded to a computer for analysis. Because the monitoring was performed under the supervision of an experimenter, the occasional bad reading was immediately replaced by a subsequent measurement. The A&D 2421 monitor has been validated extensively [21–23], and has been found to satisfy both the American Association for Medical Instrumentation and the British Hypertension Society criteria [24].

Stressor
A 3 min serial-subtraction task was used, with the subjects counting backwards, aloud, by 13s. A different starting point was used in each of the four sessions. The 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 s), 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: it is the most widely used mental task in reactivity studies [1], and 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 [25–27].

Counterbalancing
The 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
The 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, the subjects were escorted to one of the settings: inside the laboratory, a classroom located on a different floor of the same building or the subject's apartment (located within two blocks of the laboratory). The subjects were studied at approximately the same time of day in all sessions. Upon reaching the setting, the subjects were seated and instrumented using the ambulatory blood pressure monitoring arm cuff. They sat upright in a comfortable position, with hands on a table in front of them, for a 12 min resting baseline period. Blood pressure and heart rate measurements were taken at 0, 3, 6, 9, and 12 min during this period.

Next, the experimenter explained the mathematics stressor, and the subject performed the task for 3 min. During this period, blood pressure and heart rate measurements were taken at 1 and 3 min; the subject 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 one, two, and three) or the subject was paid (session four).

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 (blood pressure monitors, automated blood pressure machines, and computers). The room was approximately 3 by 3.5 m (10 by 12 feet), 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 blood pressure/heart rate measurements were taken during the resting 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 the two measurements. The change score was computed as the task level minus the baseline level.
Table 1: Blood pressure and heart rate resting levels and reactivity change scores by setting (n = 24)

<table>
<thead>
<tr>
<th>Setting</th>
<th>Systolic blood pressure</th>
<th>Diastolic blood pressure</th>
<th>Heart rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline Change</td>
<td>Baseline Change</td>
<td>Baseline Change</td>
</tr>
<tr>
<td>Laboratory 1</td>
<td>97.1 ± 11.2 12.1 ± 9.7</td>
<td>73.0 ± 4.3 63. ± 4.3</td>
<td>59.2 ± 5.3 6.5 ± 4.3</td>
</tr>
<tr>
<td>Laboratory 2</td>
<td>94.7 ± 10.0 13.7 ± 6.9</td>
<td>72.9 ± 4.8 61. ± 4.5</td>
<td>58.5 ± 4.8 6.7 ± 4.5</td>
</tr>
<tr>
<td>Classroom</td>
<td>97.8 ± 10.1 14.5 ± 7.7</td>
<td>72.7 ± 4.9 60.2 ± 4.5</td>
<td>58.4 ± 3.8 6.3 ± 5.1</td>
</tr>
<tr>
<td>Home</td>
<td>95.3 ± 11.9 14.6 ± 11.0</td>
<td>72.1 ± 4.9 67. ± 6.8</td>
<td>59.7 ± 4.9 6.5 ± 4.8</td>
</tr>
</tbody>
</table>

Values are expressed as means ± SD.

Pearson correlations were computed from the cardiovascular means and change scores taken at each of the locations. These were provided to represent the test–retest coefficients used in classical test theory, taken across laboratory sessions 1 and 2. In addition, correlations were computed between each of the laboratory and non-laboratory sessions. G coefficients were computed for the resting baseline and changes.

**Results**

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

Table 1 shows the mean cardiovascular resting levels and change scores for each setting. These values did not vary a great deal across the different settings, compared with the SD of the measurements within each setting. For example, the range among systolic blood pressure levels was less than 3 mmHg; the lowest measurement was 94.7 mmHg (laboratory 2) and the highest was 97.6 mmHg (classroom), compared with within-setting SD averaging 10.8 mmHg.

Test–retest reliability of cardiovascular levels and change scores

Pearson test–retest correlations are shown in Table 2 (baseline levels) and Table 3 (change scores), to allow a comparison between the applications of classical test theory and generalizability theory. The correlations between each of the laboratory sessions and the classroom and the home settings are also provided. Table 2 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 blood pressure (r = 0.63) and heart rate (r = 0.68). Table 2 also shows that correlations between the laboratory settings and the classroom and home settings were similar to the test–retest correlations, indicating good generalizability.

For change scores, the pattern was different. For the two blood pressure measures, the test–retest correlations were fairly strong, with r = 0.68 and 0.62, for systolic and diastolic blood pressures, respectively (Table 3). For the heart rate, however, the test–retest correlation between the change scores was almost zero.

For the blood pressure change scores, the correlations between the laboratory and the nonlaboratory sessions were, 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. However, the order was not completely counterbalanced, since the two laboratory sessions were always contiguous. This is a potential confounding factor, since the test–retest correlations were always 1 day apart, and this was not always true of the laboratory/nonlaboratory correlations. We therefore compared generalizability correlations, collapsed across measures, as a function of number of days apart. For baseline measures, the correlations were, respectively, 0.69 (1 day), 0.65 (2 days) and 0.67 (3 days). For change scores, the correlations were 0.27 (1 day), 0.31 (2 days) and 0.31 (3 days). Thus, it appears that the number of days separating sessions did not systematically affect the associations.

G coefficients among cardiovascular baseline levels

G coefficients, indexed by intraclass correlations, were computed for the baselines and change scores. The design for this analysis is that of a G study with one factor: setting.
Table 4 shows the estimated variance components, the percentage of variance accounted for by setting, and the \( G \) coefficient. This coefficient is interpretable in a manner similar to a reliability coefficient in classical test theory.

The first point of interest in Table 4 concerns the percentage of variance explained by setting. For all three measures, the value was zero, or very close to it (zero values substituted for actual estimates, as recommended by Cronbach et al. [18]). This confirms the observation from Table 1 that blood pressure levels and changes did not vary a great deal across settings.

The estimated variance components given in Table 4 provide a method of comparing the variance due to each of the components: person, setting, and the residual. Although the mean square terms derived from the analysis of variance source table are used in the computation of the estimated variance components, they are not identical to these components.

The \( G \) coefficient for systolic blood pressure was fairly high (0.82). This suggests that the mean resting level taken on one occasion (averaged, in the present case, across three measures) is highly generalizable to other settings. The estimated variance components shown in Table 4 help to show why this correlation is so strong. The correlation is based on the ratio of 'true score variance' (estimated using variance due to persons) to the total variance. Table 4 shows that the variance due to persons was far greater than either of the other estimated variance components. For example, for systolic blood pressure baseline levels, the variance due to persons was 96.44, compared with variances of 1.13 (variance due to settings) and 21.09 (residual variability). When the variance associated with persons is large relative to that associated with either the setting facet or the residual, the correlation must be large. Overall, the \( G \) coefficients for levels were similar to the test-retest correlations reported in Table 2.

Generalizability coefficients among cardiovascular change scores
As Table 4 shows, the generalizability for changes was quite different than that for levels. For blood pressure, the \( G \) coefficients for the change scores were quite poor (0.47 for systolic; 0.36, for diastolic blood pressure). For the heart rate change, the generalizability was very poor, with a coefficient of 0.13. An inspection of the estimated variance components (Table 4) illustrates this. For systolic blood pressure, for example, the variance component for the residual was about as large as that for persons (which is considered the true score variance), and so the ratio of true to total variance is small. The \( G \) coefficient, of course, is weakened by any source of variance other than that for persons.

Residual variance
As is almost always the case in analysis of variance models, some of the variance remains 'unexplained' by persons, and by the factors (or, in the \( G \) theory model, the facets). This is termed the 'residual' variance, and often appears in the analysis of variance source table as the error term. As noted earlier, the residual term comprises variance due to the person-by-setting interaction and variance due to random error; these are indistinguishable. It is crucial to be able to unconfound, and partition these, however, because the variance due to the person-by-settings interaction is at the heart of the generalizability issue. If the variance due to the persons-by-settings interaction is high, this indicates that the relative ordering of individuals is not stable from one location to the next, which is another way of saying that the response does not generalize well from one setting to the next. If, alternatively, the variance is mostly due to error, this indicates that random influences are influencing the measurements, which do not bear on the generalizability. Here, we partition the residual term into the two components: random error and the persons-by-settings interaction.

Random error and the person-by-settings interaction are ordinarily confounded, and lumped together as the residual. However, because two identical sessions were included in the experimental design (the two laboratory sessions), the variance due to 'pure error' can be estimated; the variance associated with the two laboratory sessions must be attributable to random error; no variance can be attributable to a change in setting, because setting did not vary. The variance due to the person-by-settings interaction can then be computed by subtracting the error variance from the total residual variance.

The independent estimate of error and of the person-by-setting interaction was made by calculating the sums of squares associated with the measurements taken during the two laboratory sessions, and subtracting this from the total sums of squares. The degrees of freedom were then also partitioned accordingly, and the mean square for each component was then calculated by dividing each of the sum of squares by the associated degrees of freedom. Finally, estimated variance components were computed on the basis of the partitioned mean square terms.
Table 5. Sums of squares (SS), mean square (MS), estimated variance components (EVC), and percentage of variance due to persons-by-settings interaction, for baseline levels and change scores (n = 24)

<table>
<thead>
<tr>
<th></th>
<th>Baseline levels</th>
<th>Change scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS</td>
<td>MS</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person-by-setting</td>
<td>939.27</td>
<td>20.42</td>
</tr>
<tr>
<td>Error</td>
<td>615.63</td>
<td>22.42</td>
</tr>
<tr>
<td>Diastolic blood pressure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person-by-setting</td>
<td>375.09</td>
<td>8.13</td>
</tr>
<tr>
<td>Error</td>
<td>177.47</td>
<td>7.72</td>
</tr>
<tr>
<td>Heart rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person-by-setting</td>
<td>422.85</td>
<td>8.34</td>
</tr>
<tr>
<td>Error</td>
<td>188.65</td>
<td>8.20</td>
</tr>
</tbody>
</table>

*Negative values set to zero. Degrees of freedom are 46 for the persons-by-settings term, and 23 for the error term.

Table 5 shows that, for systolic blood pressure, the person-by-settings interaction accounted for a fairly large proportion of the residual variance (44%). For diastolic blood pressure, the percentage was smaller (15%), and for heart rate the percentage fell between these (30%). These numbers are consistent with the attenuation of the Pearson correlations as the setting changes (compared with the test-retest Pearson correlations).

Attenuation of generalizability due to unreliability
In order to estimate the theoretical magnitude of the laboratory-nonlaboratory correlations, and the G coefficients, had the measurements been perfectly reliable, the Spearman correction for attenuation was applied. The test-retest correlations served as the estimates of reliability required to solve these equations. Table 6 shows the uncorrected and corrected Pearson and G coefficients. For simplicity, the averages of the Pearson correlations are shown (converted using the Fisher Z). Because of the extremely poor reliability of the heart rate changes, the corrected scores for this parameter are not shown. Table 6 indicates that, once the unreliability has been accounted for, generalizability for the blood pressure change scores attains moderate levels, with G coefficients of 0.69 for systolic blood pressure and 0.58 for diastolic blood pressure.

Table 6. Average Pearson correlations and G coefficients corrected for attenuation due to unreliability for systolic blood pressure (SBP), diastolic blood pressure (DBP), and heart rate (HR) levels and changes

<table>
<thead>
<tr>
<th></th>
<th>Baseline levels</th>
<th>Change scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SBP</td>
<td>DBP</td>
</tr>
<tr>
<td>Average Pearson r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncorrected</td>
<td>0.64</td>
<td>0.50</td>
</tr>
<tr>
<td>Corrected</td>
<td>1.00*</td>
<td>0.95</td>
</tr>
<tr>
<td>G coefficients</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncorrected</td>
<td>0.62</td>
<td>0.59</td>
</tr>
<tr>
<td>Corrected</td>
<td>1.00*</td>
<td>0.94</td>
</tr>
</tbody>
</table>

*Because the average correlation is greater than the reliability, the corrected coefficient is greater than 1, and has been set to 1.00. NR, extremely poor reliability in this measure, and the corrected coefficients will be highly unstable.

Table 6 shows the average correlations and G coefficients for baseline and change scores for systolic and diastolic blood pressure, and heart rate. The corrected coefficients are used to estimate the generalizability across a set of several conditions (four locations, in the current study). A Pearson correlation can only be used to describe bivariate relationships. The present analysis was a simple one: one factor
(location) with four facets. However, more complex
designs can also be studied, in which several factors are
studied simultaneously. For example, generalizability can
be studied across several locations (as in the present
study), across different races and genders, using different
measuring techniques, different experimenters, and so on.
The ability to examine these factors simultaneously allows
the interpretation of interactions between two or more
factors; again, this cannot be accomplished using tradi-
tional correlational methods.

Resting levels
The test-retest correlation was quite strong for systolic
blood pressure baseline levels, whereas the correlations
were somewhat weaker for diastolic blood pressure and
for the heart rate. When examined across all four settings,
the G coefficients for levels were similar to the test-retest
reliability scores for each parameter, indicating that the
change in location did not affect blood pressure or heart
rate levels.

The generalizability for systolic blood pressure was
quite high, approximately 0.8, which suggests that systolic
blood pressure measures taken in the laboratory or clinic
(and averaged across three individual measures, in
the present case) are strongly representative of measures
taken in nonclinic or nonlaboratory settings. The G coef-
ficients for diastolic blood pressure and heart rate levels
were somewhat poorer than for systolic blood pressure,
but were still in the moderate-to-strong range (with G
coefficients of approximately 0.6 for both parameters).
These results parallel those of other studies, in which
blood pressure and heart rate levels taken in the clinic or
laboratory served as useful predictors of levels measured
in the natural environment [8].

Change scores
For the blood pressure measurements, the test-retest
correlations were in the moderate range, with r = 0.68 and
0.62, for systolic and diastolic blood pressures, respec-
tively. However, the G coefficients between the labora-
tory and nonlaboratory settings were smaller (Table 4).
For heart rate, the test-retest correlation was close to zero,
and the G coefficient was approximately the same as
the test-retest association. Taken together, the Pearson
correlations and the G coefficients suggest that a change
in setting, even with all other aspects of the test situ-
ation held constant, affects the cardiovascular stressor
responses.

The interpretation of these coefficients requires an
inspection of the estimated variance components
(Table 4). The G coefficient for systolic blood pressure
change (0.47), for example, was quite low, in terms of
traditional psychometric practices. Table 4, however,
shows that the poor generalizability was not due to a main
effect of setting; for all three measures, the setting facet
accounted for none of the explained variance. Instead, the
G coefficient was poor due to the residual variance, which
comprises variance due to two sources: error and the
person-by-settings interaction.

The poor generalizability of change scores was due to
either or both the sources of variance included in the
residual term. That is, the small G coefficients were due
to the error component (which is an aspect of the measure-
ment process, but provides no information about general-
izability), or to the persons-by-settings interaction.
The latter does reflect upon the generalizability: variance
associated with the interaction indicates that some indi-
viduals are more reactive in some settings than in others
and that the settings which produce greater reactivity
are different for some individuals than for others. As
described earlier, these two sources of variance, error and
the person-by-settings interaction, usually cannot be
disentangled in the residual term.

Because two identical levels of the settings facet were
included in the study, however, we were able to further
partition the residual variance into the two separate
components. As Table 5 shows, the interaction was most
prominent for systolic blood pressure changes, comprising
44% of the residual, compared with only 15% for diastolic
blood pressure.

In summary, the data suggest that the generalizability of
the change scores was diminished due to the person-by-
settings interaction, especially for systolic blood pressure
changes. After correcting for unreliability, the G coeffi-
cients become somewhat stronger. Even the corrected
coefficients, however, indicate that most of the variance
in the non-laboratory change scores remains unexplained
by the laboratory measurements. It seems clear that the
poor generalizability of the change scores is not simply
due to unreliability of the measurements. In practice, the
relationships will of course be even weaker, since the
observed measurements cannot be corrected for unreli-
ability. It appears that the simple change of setting is
capable of producing changes in reactivity, with different
individuals showing greater or lesser reactivity in different
settings.

Conclusions
From the point of view of the clinician or researcher, it
is convenient and efficient to be able to take measure-
ments either in the clinical or in the laboratory setting as
a means of estimating the individual's characteristic blood
pressure levels or stress responses. For the clinician who
is primarily interested in taking casual pressures as a
means of assessing the risk of cardiovascular disease, the
levels measured in the clinic setting appear to generalize
quite well for most individuals outside the clinic setting.
This has been demonstrated in other studies, in which
mean resting levels were strong predictors of mean levels
taken using ambulatory monitoring techniques [8]. In the present study, as well, the data indicated that a change of setting did not appreciably affect levels. However, there is at least one population in which clinic measurements will lead to biased estimates: those patients diagnosed as having white-coat hypertension. This is a minor concern: Pickering et al. [29] have estimated that approximately 20% of patients diagnosed with essential hypertension are hypertensive only when being examined in the clinic. In contrast, their blood pressures are much lower, and may fall within the normal range, when measured outside the clinic, even at work (which in general produces some of the highest blood pressures that will occur during the patient's day).

For researchers and clinicians who are interested in stress responses, or reactivity, measurements taken in the laboratory do not appear to generalize very well to non-laboratory settings, even when all parameters except location are maintained constant. 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 seems likely that the specific challenges and stressors actually encountered represent the most significant changes between the laboratory and the natural environment. Laboratory stressors are selected, among other reasons, in order 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 moderately generalizable to non-laboratory situations.

In order to learn to what precise domains we can legitimately generalize, factors which may be sources of variability must be investigated. This is crucial information for the design of future studies, because those factors which do not matter much (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 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 little affected by different times of day, i.e. morning versus afternoon. However, other measures, such as cortisol, are heavily influenced by the time of day during which testing occurs [30]. Thus, 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, is crucial to the design of reactivity studies.

The present study demonstrates that, to find predictive power extending from the laboratory to the natural environment, there is no dimension of variability so trivial that it can be dismissed without investigation. This is not so much a concern when the focus is on resting levels, or casual blood pressures, which appear to remain stable under a variety of circumstances. However, for stress responses, it is a great concern. If simply changing the location of the test site can reduce the laboratory–life associations, 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.

References


