Aggregation Bias and the Use of Regression in Evaluating Models of Human Performance

NEFF WALKER[1] and RICHARD CATRAMBONE, Georgia Institute of Technology, Atlanta, Georgia

Regression analyses are increasingly being used to provide confirmatory evidence for models of human performance. The amount of information made available to judge these models is reduced because clearly established standards in the techniques of performing and reporting regression analyses are lacking. This paper addresses two primary problems in regression analysis: aggregation of data and the aggregation of variables into composite models. We provide examples of the misuse of regression techniques and recommend ways in which the amount of information made available to evaluate the model being tested can be maximized in analysis and reporting.

REGRESSION MODELS AS SPECIFICATION SEARCHES

In the last ten years there has been a shift away from domain-specific psychological theories toward large-scale models of human performance. These models—e.g., Adaptive Control of Thought [ACT*], Anderson, 1983; Cognitive Complexity Theory, Kieras and Polson, 1985; Goals, Operators, Methods, and Selection [GOMS], Card, Moran, and Newell, 1983; State Operator and Results [SOAR], Laird, Newell, and Rosenbloom, 1987—seek to identify the cognitive and, sometimes, the perceptual and motor processes involved in performing a task and to specify how these processes interact. The models are often used to explain performance on various computer-based tasks, partly because such tasks are constrained and therefore more easily quantifiable. Reliance on these models has led researchers away from using factorial designs tested through analysis of variance and toward task-performance comparisons that use regression analysis to build and verify the model.

In this paper we will suggest that, in the process of building these models of performance, researchers have failed to establish consistent standards for the use and reporting of regression analysis. Two frequently used practices—aggregation of data and aggregation of variables into composite models—cause biases in the building and evaluation of performance models and reduce the amount of information made available to judge the models.

Leamer (1978) has described the primary purpose of regression analysis in research as a process of "specification search." Specification search is the attempt to build a model of

performance and to provide information about the model's completeness (level of specification). In a specification search, regression is used to uncover variables relevant to performance and to determine how these variables interact to determine performance.

Regression analysis can be a powerful tool, yet it is not without drawbacks. Numerous articles have been written addressing such possible misuses of the technique as violations of assumptions, collinearity, and effects of sample size (e.g., Bibby, 1977; Carpenter, 1984; Gordon, 1968).

Our paper neither directly addresses any of these issues nor attempts to provide mathematical proofs for the issues we raise, although appropriate references are provided. Instead, we borrow our approach from Mulaik's work on factor analysis (Mulaik, 1991), in which he suggests that factor analysis should be viewed as a practice, not merely an algorithm. We will argue that the primary purpose of the practice of regression should be to maximize the amount of information made available to build and evaluate models of performance, not to find the highest possible value of R^2.

In regression analysis, the primary measures are the regression estimates (variable weightings and intercept value) and the proportion of explained variance (R^2). The regression estimates are used primarily to make predictions and to evaluate the relative importance of the variables. Once one knows the estimates of the intercept and the coefficients for the variables in the equation, these values can be used to make predictions about performance in a new task. The value of R^2 is important because it provides a rough measure of the level of specification of the model. A fully specified model would yield an R^2 value of 1.0. When the value of R^2 is high, one often assumes that the model tested provides an adequate theoretical explanation of the task under study. When the value is not high,

one assumes that important variables are missing from the model, and the theory is judged to be incomplete. In this paper we will present and discuss practices that affect the regression estimates and the value of R^2 yielded by regression analyses and how some of these practices can make it difficult to evaluate the level of specification of the model.

A second important aspect of evaluating a model is the model's ability to explain a large percentage of variance in performance, as measured by R^2, and to "simply" do so. For example, one model might have many predictor variables, and these variables might require a great deal of effort to operationalize (for instance, they might require a complicated production system analysis). This model explains 75% of the variance in performance. Suppose another model has two predictor variables (e.g., number of keystrokes and number of letters on the screen), both of which are simple to operationalize and measure. A regression analysis using this model explains 70% of the variance in performance on the same task. Which model is better? If the purpose is to simply maximize the percentage of variance explained, then the first model is better. Yet the second provides almost as good a description of performance with fewer variables and considerably less effort.

If the model is being used to evaluate a new interface design, the answer might depend on the ease with which one can apply the two models. In this case the second model would undoubtedly be the preferred model. Model simplicity has practical importance and, as we will show, sets certain requirements about how regression analysis should be used and reported. We are not advocating that all models must be abandoned if they are complex or require effort to apply. However, because of the misuse of regression techniques, the theoretical status of many variables in complex performance models may not be as

strong as they appear. The inappropriate inclusion of such variables in the model limits the model's theoretical value and practical application.

This paper has three primary goals. First, we will identify two broad sets of practices in the use and reporting of regression analyses and discuss the limitations of these practices. These practices are the aggregation of data (aggregating over ranges of values, trials, or subjects) and the use of composite variables in a model (failure to test alternative models, prespecifying relationships among variables). For each practice we will present examples from human-computer interaction (HCI) in which researchers have used one of these practices to validate a model and discuss the effects of these practices on model evaluation. In the final section of the paper we will recommend techniques for the use and reporting of regression analysis that can eliminate these problems.

AGGREGATION OF DATA

In some regression analyses, mean values of the independent variables are used to predict mean values of the outcome variable. This practice has been referred to as *aggregation* (e.g., Klein, 1946). Many papers in sociology and economics have focused on the methodological issues arising from the use of aggregated data (see Doggan and Rokkan, 1969, for a review of some of this work). In these fields researchers in the past often had to rely on secondary data sources, which provided only aggregated data. For example, researchers using census reports and tax-base information had access only to data for ranges of incomes, not for the individual households. When building models of spending patterns of households, the researchers had no alternative but to use aggregated data in their regression analyses. In current work in human performance, aggregation over

data is a choice made by the researcher. This choice, however, invites error.

To what extent can findings based on a model derived from aggregated data be applied to individual behavior? In the census data example, can one use a model based on average or grouped performance to draw conclusions that are valid for individual households? The problems associated with drawing inferences about individual performance from results of regression analysis using aggregated data have been referred to as the *ecological fallacy* (e.g., Robinson, 1950) and *cross-level analysis* (Hannan, 1971).

As Knight (1984) noted, aggregating over variables amounts to setting the partial regression coefficients for these variables to zero. This assumes that the variables have no relationship to performance—or, at least, no effect on it that is of interest to one's understanding of task performance. Aggregating over data or variables can restrict the cases to which the model's explanatory power applies. This is especially true when the regression analysis aggregates over a range of values of the predictor variable, thereby eliminating a source of variation that should be explained by the model.

Types of Aggregation

Aggregation is the process of collapsing over data points and then using a measure of central tendency to represent that range of data. When evaluating the effects of aggregation, one must determine what was different about the data points that were aggregated, or, in Knight's (1984) terms, which partial coefficients were set to zero. For instance, in the census example the data were points that represented household income. In the regression using household income as a predictor variable, variance in that variable had been removed.

Aggregation can be used in three primary ways: aggregation over trials or repeated

measures, aggregation over subjects, and aggregation over ranges of the predictor variables. The purpose and possible effects of each type of aggregation are discussed next.

Aggregation over trials. One common practice is to aggregate multiple trials with the same levels of the predictor variables to create a mean measure of performance. The rationale for this type of aggregation is that sensitive dependent measures, such as reaction times, can be badly skewed by differences in attention, learning over trials, or other factors. The effect of these extraneous variables is reduced by collapsing over the multiple observations, which results in means that are believed to provide a closer approximation to the true value. By aggregating over multiple observations of the same task, error variance is removed from the equation.

Aggregation over subjects. The purpose of this type of aggregation is to remove the effects of differences in subjects' abilities on performance by eliminating the variance associated with the individual differences through the use of mean group performance rather than individual scores. In this type of aggregation, the removed variance might be assumed to be unrelated to the predictor variables in the model if individual differences do not interact with the treatment variables.

Aggregation over values of a predictor variable. The third type of aggregation is similar to the previous census data example. Here data are aggregated across the values of a predictor variable in the model. For example, if household income is hypothesized to be a predictor of the proportion of total spending on transportation, and categorized ranges of incomes (e.g., $10,000 to $20,000) are used as the predictor variable, then one is aggregating over levels of the predictor variable. This type of aggregation, which we will refer to as *variable aggregation,* removes variance that is directly related to the predictor variable—in this example, family income.

Dangers of Aggregation

Reductions in the amount of variance in the data through aggregation can both eliminate important information and falsely inflate the value of R^2 yielded by the regression analysis. Although these two issues are related, we will address them separately.

Inflating the value of R^2. Cramer's (1964) work is helpful in determining the amount of inflation in the value of R^2 that has been introduced by the aggregation. As Cramer and numerous others have suggested (e.g., Haitovsky, 1973; Pearson and Hartley, 1958; Prais and Aitchison, 1954; Thorndike, 1939), the use of grouped data forces aggregation across a source of variation. If there were no variation within the values (or variables), the effect of grouping on R^2 would be nil. This is rarely the case. When there is variation in these values, regression analysis yields an inflated value of R^2. The amount of inflation cannot be determined precisely unless one has access to the individual data points. However, Cramer developed a two-step estimation procedure (based in part on the work of Pearson and Hartley, 1958) that can be used to estimate the magnitude of bias resulting from aggregation in the value of R^2.

First, R^2 is transformed into a C value using the equation:

$$C = R^2/(1 - R^2). \qquad (1)$$

This transformation changes the range of values for the measure of explained variance. C can now range from 0 to infinity.

Second, C is entered into the following equation:

$$C^*/C = (t - 2)/(N - 2). \qquad (2)$$

In this equation C^* is the transformed R^2 for the grouped data, and C is the R^2 for the ungrouped data. N is the total number of observations or values of the predictor variable,

and t is the number of groups or means used in the grouped data regression.

To see how this estimation procedure works, take a case in which 60 observations of a variable have been aggregated into six ranges. For this example we will assume aggregation over subjects and trials without considering their possible effects. The means for these six ranges are regressed on an outcome variable, and the resulting regression yields an R^2 of 0.90. Applying the transformation of R^2 (Equation 1), the C value is 9.0. Using this value in the estimation equation (2) yields a value of $C^* = 0.62$.

Using the value of C^* in place of C in Equation 1 produces an estimated value of R^2 for the ungrouped data of 0.38. Because R^2 is often used to judge how well the model specifies the component processes of performance, the R^2 from the grouped data suggests an almost complete specification. The estimated R^2 for the ungrouped data clearly does not. Of course, these adjusted values of R^2 are only estimates. The actual degree of inflation in the value of R^2 caused by the use of aggregated data depends on the nature of the variation in the aggregated variable (e.g., are the data equally distributed across the range of values being aggregated?).

Type of information eliminated. The second issue that must be addressed to determine the effects of aggregation on the evaluation of the model tested by regression relates to the type of information that was eliminated by aggregation. Here one must determine which source of variance was eliminated and how that relates to the model being tested. For example, when one collapses over values of a predictor variable, one is reducing, at least in part, variance that is actually caused by the variation in the predictor variable. Given that different values of this variable almost certainly cause variation in the outcome variable, one is strongly biasing the test of the model. As many researchers have shown (e.g.,

Haitovsky, 1973; Pearson and Hartley, 1958; Prais and Aitchison, 1954; Thorndike, 1939), the bias results in an inflated value of R^2.

Aggregating over repeated measures (multiple observations) of a single value of the predictor variable or over subjects can also bias the value of R^2. However, the effects of these biases on interpretation of the level of specificity of the model might not be as severe as those caused by aggregation over ranges of the predictor variable because the variance being removed from the data is not directly related to the model. Collapsing over repeated measures assumes that the variance lost is not related to the value of the predictor variable but is, instead, at least partly a result of error. When aggregating over scores from different subjects, one is removing variance arising primarily from individual differences in background, motivation, or abilities. If this source of variation is beyond the scope of the model, then there is no problem associated with aggregation across subjects.

An Example of Aggregation in HCI

John and her colleagues used a variant of the Goals, Operators, Methods, and Selection model to explain stimulus-response compatibility (e.g., John and Newell, 1987; John, Rosenbloom, and Newell, 1985). They created algorithms of the perceptual, cognitive, and motor processes required to recall and use abbreviations of computer commands. Their resulting model was then used to predict the time required to recall and use computer command key bindings.

Based on their model, John and Newell (1987) predicted the performance times for three sets of key bindings for 12 computer commands (e.g., $^\wedge$D for "delete"). A set of possible algorithms was created for each command in each condition that modeled the processes required by a person to produce the appropriate key binding. The algorithms for

each command included four types of operators (perceptual, mapping, retrieval, and motor). With the operators specified for each command, all that was needed to describe performance was a set of values (times) required for each of these operations to be performed.

John and Newell chose values based on their earlier work (drawn from the regression analyses presented in John et al., 1985). With this information they could calculate predicted times for both the time to the initial response (hitting the first key) and execution time (time from the first key press until the response was completed) for each command in each of the three key binding conditions. This allowed John and Newell to predict the time required to produce each key binding and to compare these predictions with performance times. Using the six predicted times (the mean predicted times for the time to the first key press and the times for the remaining key presses for each of the three key binding systems) as the predictor variable and six mean performance times as the outcome variable in a regression analysis, they found a value of $R^2 = 0.776$.

Although this R^2 certainly shows that this model can be used to predict average performance on this task, does it provide evidence that the model was well specified? This question can be addressed by first estimating the value of R^2 after correcting for the various types of aggregation used in the analysis and then determining what data were aggregated.

In the analysis reported by John and Newell (1987), six predicted times were regressed on six mean performance times (three each for initial response times and execution times). To obtain these six predicted times and the six observed means, John and Newell aggregated over repeated trials (five trials for each command in three key binding systems), over subjects ($N = 12$), and over the commands used in each key binding system (12).

Collapsing over repeated observations of the same command seems to be a valid practice because this analysis is using response times as a dependent measure. Aggregating across subjects might also be a valid practice if John and Newell (1987) did not intend to explain individual performance. However, aggregating over the different commands in a system of key bindings is more problematic.

Each of the three systems of key bindings used 12 different commands. These commands yielded different values for the four operators in John and Newell's model. All of these values were then aggregated to produce four mean values for the operators for each key binding system. The operators were then multiplied by the weightings found in previous work (John et al., 1985) and summed to produce the predicted times for performance. This aggregation eliminated variance directly caused by variation in the predictor variables of the model.

A reanalysis. Cramer's (1964) correction equations can be applied to estimate the effects of aggregating over the values of the four operators in John and Newell's (1987) model. Using Equation 1, the R^2 reported by John and Newell produces $C = 3.464$. The value for the right side of Equation 2, $(t - 2)/(N - 2)$, becomes 0.057, resulting in a C^* of 0.198. From these calculations the estimated value of the R^2 for the ungrouped data is 0.17. If this estimate of the R^2 of the ungrouped data is correct, it raises serious questions about the level of specification of their model.

Of course, the estimated value of R^2 (0.17) is only an estimate. John and Newell's (1987) analysis may yield an R^2 value greater than this estimate if performed on the unaggregated data. The accuracy of Cramer's (1964) estimate depends on the amount of variance removed from the model by the aggregation. If no variance was lost (meaning that times for all initial and concluding keystrokes within conditions were identical when

collapsed over subjects), then 0.776 would be the true value of R^2. If variance was eliminated from the analysis, then the value of R^2 on the unaggregated data would be lower. Based on the reported data, however, we can only conclude that the possible value of R^2 ranges between 0.17 and 0.776. Given the breadth of this range, the reader cannot make an independent evaluation of the level of specification of John and Newell's model.

THE USE OF COMPOSITE VARIABLES

A second common practice in regression analysis that can limit the amount of information about the model being tested is the use of composite variables or composite models. This issue has received considerable attention in the areas of tests and measurement and individual differences (e.g., Paunonen and Gardner, 1990; Perloff and Persons, 1988). In these areas composite measures are often used to provide a test of a single factor or dimension. This practice is sometimes referred to as *aggregation*, but in this paper a distinction is being made between aggregating over ranges of a variable or data, as outlined earlier, and the aggregation of multiple predictor variables to create a single predictor variable, which we will call a *composite variable*.

An example of the use of a composite measure might be a variable that consists of subscales or multiple measures that have been combined to create a score of intelligence. The problems associated with the use of composite variables or a composite model (in which all the variables are forced into the regression in a block) as a single predictor stem from the possible intercorrelations among the constituent (subsumed) variables. In general, subscales or multiple measures of the same variable or factor are assumed to increase the reliability of the composite variable score. However, when the subscales are not multiple measures of the same factor or

variable, combining them into a single composite measure or index may be inappropriate. Although important, this specific issue has been discussed at length elsewhere (cf. Perloff and Persons, 1988; Rushton, Brainerd, and Pressley, 1983) and will not be addressed extensively here.

Composite Variables Restrict Possible Relationships That Can Be Considered for the Subsumed Variables

The key issue associated with using composite variables is that their use restricts the ways in which subsumed variables may be related to one another. For example, if a test of intelligence has eight subscales and combines the subscale scores to create an index of intelligence, it is assumed that the subscales have correct (often equal) weightings and that they do not interact in a way that relates to performance. One will not know if this is not the case unless the other possible relationships among the subsumed variables are investigated and reported. Thus, when creating a composite variable, one has limited the types of models that can be tested in a regression by (perhaps inadvertently) prespecifying the relationship that exists among the subsumed variables.

The problem associated with using a composite variable is that there might be tasks in which the relationship among the subsumed variables changes. For example, there could be a task in which high performance on a subscale is inversely related to performance, whereas other subscales have a direct relationship to performance. Using the prespecified composite variable would result in a lower overall level of association than if the relationship among the subscales had not been prespecified.

When one uses a composite variable or forces a group of variables to enter into the regression concurrently (what we are referring to as a composite model), one cannot

determine the relative importance of the subsumed variables. This reduces the amount of information that can be gained from the regression analysis.

The problems associated with the use of composite variables/models can be broken into at least three broad areas. The first problem area is that researchers who test a composite variable often fail to test the individual importance of the component variables in the model. Second, by using a composite variable in a model, the researchers are, often unknowingly, prespecifying the type of relationship that can exist among the variables. Other models, with different types of relationships among the subsumed variables, are not tested. This does not mean that one should never use a composite model or variable. There are areas in which we have strong empirical research to support the specific set of prespecified relationships that are formalized in the composite variable. However, as we will show, even in such cases, testing alternative models could be beneficial. Finally, researchers using composite variables or models can analyze their data appropriately but can fail to report all of the analyses performed. This, of course, can occur with any type of analysis, but it can be especially problematic in the use of regression. Each of these problems will be discussed next.

Failure to Test Subsumed Variables in a Composite Variable

An example of the use of a composite model to predict performance is, again, found in the work of John and Newell (1987). Their analyses did not allow the reader to determine the relative importance of the variables subsumed by the composite predictor variable. Recall that John and Newell used a single value as their predictor variable for both the time to the first keystroke and the predicted time for the rest of the keystrokes for each of

the three sets of key bindings. These predictor values were based on the average number of each of the four operators multiplied by the value found for that operator in John and Newell's earlier work. One or more of the subsumed operators could have provided most of the model's explanatory power. Fortunately, John and Newell's report did include information about the mean values of the subsumed variables. This information, plus means for the three conditions, allowed us to reanalyze their data and test the possibility that one or more of the subsumed variables provided most of the explanatory power of their model.

In an attempt to test the relative importance of the subsumed variables in John and Newell's (1987) model, we performed a regression that used only two of the four operators in their model. We selected the motor and perceptual operators because these are easily measured variables. The weightings for the motor operator (time to press a key) and the perceptual operator were multiplied by the average number of key presses and perceptual operators reported for each condition. The six resulting values were then regressed on the reported observed values (using the same aggregation procedures as used in John and Newell, 1987). This regression yielded an R^2 of 0.72. The regression using all four operators of the model yielded an R^2 of 0.776.

This analysis suggests that the two cognitive operators (mapping and retrieval) in the model might not be needed to describe performance. This has important theoretical and practical implications for an evaluation of the explanation of stimulus-response compatibility advanced by John and Newell (1987; John et al., 1985). If one's purpose is to use the model as a way to evaluate different key binding systems, should one produce more than 100 pages of production rule analyses for three 12-command systems to increase the

amount of predicted variance from 72% to 78%? The reanalysis suggests that the cognitive aspects of the model provide little explanatory power. Our analysis required little more than simply counting the number of keystrokes for each key binding condition and yielded an R^2 value almost as large as that of John and Newell's (1987).

By using a composite variable and not reporting the separate effects of the subsumed variables, John and Newell (1987) inadvertently restricted the amount of information available to the reader to judge the validity of the claims of their model. It may be that the mapping and retrieval operators do provide a large amount of explanatory power, but from the analysis John and Newell presented, one cannot make that judgment. The casual reader might assume that all four operators in their model are equally important in predicting performance. Our reanalysis suggests that their model (requiring a complex and lengthy analysis) might provide no more explanatory power than simply knowing the number of keystrokes required and whether or not the person looked at the screen.

Perhaps the best summary of our position can be made by quoting from Robert Sternberg's (1986) discussion of the use of multiple regression:

> Perhaps it is a matter of taste, but I have never been impressed by multiple regressions that contain one statistically significant independent variable and a whole bunch of statistically trivial independent variables along with it. The multiple correlation adds nothing to the simple correlation except junk—and a more impressive looking coefficient. (p. 278)

We feel that the point he makes applies to the use of composite variables. Reports of regression analysis using composite variables or models must also provide information that specifies the relative contribution of each subsumed variable.

Presupposing Relationships among Variables in the Model

A second drawback to the use of composite variables or models is that it limits the ability to test alternative models of performance. As outlined earlier, a composite variable specifies one set of relationships that may exist among the subsumed variables. This practice thereby eliminates from consideration other possible combinations of the variables. It can be especially misleading if the regression analysis also uses data that have been aggregated to the levels of the composite variable. To demonstrate how this practice can eliminate the perceived need to test other models, we will review some motor movement research related to cursor positioning with a mouse.

Fitts' law has served as the theoretical base for research investigating cursor positioning time in the area of HCI (e.g., Card, English, and Burr, 1978; Epps, 1986; Walker and Smelcer, 1990). According to Fitts' law, two independent variables—distance and target width—can be used to describe movement time. The variables of distance and target width are combined into the so-called index of difficulty, which usually takes the form of $\log_2(2D/W)$ or $\log_2(D/W + 0.5)$. This index provides a highly accurate description of the time required to make a movement, with R^2 values often above 0.90.

In HCI research, Fitts' law has been found to provide an adequate description of the time required to position a cursor with a mouse (e.g., Card et al., 1978). However, recent work has suggested that other equations may provide competing or even better descriptions of cursor positioning time (e.g., Epps, 1986; MacKenzie, 1992; Walker, Meyer, and Smelcer, 1993). MacKenzie's review of this work raises several interesting issues about how the use of a composite variable, in the form of the index of difficulty, might have

hidden underlying relations among the subsumed variables of distance, target width, and target height in determining cursor positioning time. In particular, MacKenzie (1992) reanalyzed the work reported by Card et al. (1978) and suggested that the regression analyses reported are incorrect because of the authors' failure to adjust the value of target width when they varied approach angle.

To illustrate MacKenzie's argument, we will provide a brief explanation of the methods used in the work of Card et al. (1978). In their experiment subjects were required to more the cursor from a starting location to a target area on the screen. The independent variables were distance to the highlighted text, angle of approach to the text (direction), and width of the highlighted text. Distance and angle of approach to the target area were varied by manipulating the location of the highlighted text. Target width was manipulated by varying the amount of text highlighted. Unfortunately, when angle of approach varies, the target width (the size of the target area as measured along the dimension of the primary direction of motion) also varies. The regression reported by Card et al. (1978) did not take this into account and thus used incorrect values for target width for calculating the index of difficulty.

While reviewing MacKenzie's (1992) work, we were surprised that Card et al. (1978) had found such a large value of R^2, given that the wrong value of width was used in the equation. We now realize that by using a composite variable (the index of difficulty) and aggregating over values of width and distance that yielded the same index of difficulty, Card and his colleagues eliminated much of the information that could have been used to judge the level of specification provided by Fitts' law.

In order to demonstrate how aggregating data to the levels of the composite variable affected the outcome of the regression analysis, it is important to determine the source of the numbers Card et al. (1978) used in their analysis. In their study there were five values of movement distance, four values of target width, and three values of angle of approach. These were all within-subjects variables, yielding 60 unique Distance × Target Width × Angle of Approach combinations for each subject. The 12 means used in the regression were based on the following types of aggregation.

First, angle of approach was eliminated because an analysis of variance had revealed no significant difference in positioning time attributed to angle. That left 20 unique Distance × Width combinations. Of these 20 values of distance and width, some yielded identical index of difficulty values. For example, in a study in which movement distances are 6 and 12 cm and target widths are 1 and 2 cm, the index of difficulty values would be the same for movements when distance was 6 and target width was 1 and for movements when distance was 12 and target width was 2. In the Card et al. (1978) study mean movement times used in the regression were based on all movements with the same index of difficulty. This resulted in 12 means, one for each level of the index of difficulty. Clearly, some variance might have been eliminated by this aggregation. If variance was removed, it resulted from variation in movement distance and target width. This would inflate the value of R^2. The question remains as to how much variance was eliminated.

Although the Card et al. (1978) data cannot be completely reanalyzed because all of the data points are not reported, the effects of aggregation and the use of a composite variable can be demonstrated in two ways. First, to estimate the amount of variance that may have been eliminated through the aggregation procedures used to reach the 12 values of the index of difficulty, we will refer to some of our own work on cursor positioning with a mouse (Walker et al., 1993).

This study investigated the time required to position a cursor with a mouse. Four values of movement distance and four values of target width were used as independent variables. In some cases different combinations of distance and target width yielded the same values for the index of difficulty, yet the positioning times for the different combinations differed by as much as 10% of the total movement time. Considerable variance could therefore have been eliminated by the aggregation procedure used by Card et al. (1978). If the aggregation had not been performed, the resulting lower value of R^2 would stimulate a search for alternative models to explain performance.

The use of aggregated data based on a composite model also masked other possible relationships between distance and target width that might have provided a better description of positioning time. Although there were few alternative models for positioning time when Card et al. (1978) performed their experiment, recent research has provided alternative models to Fitts' law that provide better descriptions of movement time, yielding R^2 values of above 0.90 on aggregated data. These models include power functions (Epps, 1986), distance alone (Gan and Hoffmann, 1988), or the square root of distance divided by width (Walker et al., 1993).

Different aggregation procedures might show an entirely different pattern of results if applied to the data of Card et al. (1978). To test this possibility, mean positioning time for the five movement distances was extrapolated from Figure 1 in the Card et al. report. This yielded means that are based on data aggregated on the five values of movement distance. The values for distance of movement were then used in a regression analysis to predict mean cursor positioning time. This analysis yielded an R^2 value of 0.95. Given the large body of evidence demonstrating that target width plays a role in determining

movement time, there is little chance that this value of R^2 would be interpreted as proving that distance alone determines cursor positioning time with a mouse. However, it does demonstrate that the combined use of a composite variable and aggregation eliminates variance that might be important in understanding the true causes of performance. This may, in turn, lead to false confidence in the level of specification provided by the model. This false confidence reduces the probability that alternative models of performance will be investigated.

Underreporting of Regression Analyses

The final problem associated with the use of composite variables or models in regression analysis is the failure to provide complete reporting of the analyses. An example from HCI of the use of a composite model is the transfer analysis reported by Lee and Polson (1989). The authors tested predictions about transfer of training using tasks involving an oscilloscope, based on an analysis using Cognitive Complexity Theory (Kieras and Polson, 1985). In this experiment subjects had to learn and use a set of procedures. Each procedure was analyzed to determine the number of production rules the subject had to know to perform the procedure. Some rules were required by more than one procedure.

The experimenters manipulated the order in which subjects learned the different procedures. Thus the number of new rules to be learned for a given procedure varied depending on the procedures already learned. The authors manipulated the total number of rules required for a task as well as the number of new rules required (rules that had not been learned from a previous procedure). They predicted that the number of rules and the number of new rules would predict learning time for a procedure.

To test this, Lee and Polson used a model with three predictor variables. The first was

each subject's mean performance time across all procedures, which was used to control for subject differences. The second variable was the total number of rules required to perform the procedure. The third variable in the regression was the number of new rules required to perform the procedure. These three variables were used to predict individual performance time for procedures.

No bias occurred in this analysis, which can be attributed to variable aggregation. The authors clearly identified the variables used, and they provided the number of observations used in the regression. However, they reported only that the "above regression model accounted for 77% of the variance of the 343 individual training times" (Lee and Polson, 1989, p. 119). They did not report the relative contributions of the three predictor variables in the regression, though they did report the weighting for number of rules. This raises the issue of the relative importance of the three variables in the amount of variance explained. This is troubling because subject differences, as represented by subjects' means in their regression analysis, could account for a large percentage of the explained variance. Lee and Polson did provide further information to help readers assess the relative importance of the three variables, but we found it difficult to make a determination of the relative importance of the variables in their model. This problem could have been eliminated had the authors provided simple and partial correlation coefficients for each of the variables or a full stepwise regression table. Either would have allowed readers to judge the contributions of each variable.

Researchers must remember, when using a regression analysis with a composite variable or a composite model, that the purpose of their analysis is to maximize the information provided to the scientific community. Composite variables might be appropriate when there is a large body of research to support the specific relationship embodied in the composite model. However, as the discussion of Card et al.'s (1978) work has shown, even in cases in which a strong reason exists to presuppose a particular relationship among the subsumed variables, the use of a composite variable still reduces the amount of information that might be provided by the regression analysis. Furthermore, when one aggregates the data to fit levels of the composite variable, one inflates the R^2 by eliminating variance that should be included in the regression and thereby discourages investigation of alternative models of performance.

CONCLUSIONS AND RECOMMENDATIONS

The purpose of this paper was to highlight practices in the use and reporting of regression analysis that reduce the amount of information conveyed. We have argued that aggregation of data and the aggregation of variables into composite variables and models can bias the outcome of the regression analysis and reduce the amount of information made available. Both practices may, however, have appropriate uses. The challenge is to determine when these practices are appropriate.

We believe that the confusion about how to use and report regression analyses stems from two sources. First, there is no clearly established reporting procedure for the use of regression, as there is for other statistical techniques (e.g., analysis of variance). Rough guidelines exist but are usually not specific. Second, regression analysis is a more flexible statistical technique, as compared with analysis of variance, allowing for multiple analytic approaches to a set of data. Therefore no simple, algorithmic rules can define how all regression analyses should best be performed or reported.

To guide our recommendations we would like to return to Mulaik's (1991) suggestion

that factor analysis and, in our opinion, regression analysis are best viewed as a set of practices, not as an algorithm. The goal of regression analysis should not be to find an answer but, rather, to maximize the amount of information made available. We hope that we have demonstrated that both aggregation and forcing variables into composite models without testing or reporting the contribution of individual variables are information-reducing techniques.

Recommendations for the Use of Aggregation

When deciding whether or not to aggregate, remember that aggregation always reduces variance. One must decide if the variance being eliminated should be explained by the performance model under investigation. We believe that aggregation might be appropriate over repeated observations of the same conditions. This practice primarily eliminates variance resulting from error and learning effects, which can reasonably be eliminated from consideration if one is interested only in asymptotic performance. However, even in such cases, aggregation over trials also eliminates variance that could be related to learning or skill acquisition, or the interaction of learning and the other variables, and therefore must be used with caution.

The common practice of aggregation over subjects is more problematic. Our initial position was that this practice is never appropriate. However, there might be exceptions in cases in which researchers are comparing their value of R^2 with that of other researchers who used aggregated data, or if the goal is only to describe average performance. Research in the area of motor movement has frequently used data aggregated over subjects. The practice is now so widely accepted that if one were to try to publish a study today that investigated Fitts' law and reported an R^2 value of 0.62 between the index of dif-

ficulty and movement time obtained through analysis of *un*aggregated data, most researchers would interpret the unusually low R^2 value as evidence that Fitts' law did not apply to the task, or that the researcher had made an error. Yet it seems that this use of aggregation over subjects is based more on tradition than on sound research practice.

There are two primary arguments against aggregation even when it is the norm. First, individual subject differences often account for more than 50% of the variance in performance in many tasks (Egan, 1988). The most appropriate analysis of these data would therefore *not* aggregate over subjects. Information about individual differences must be available in order to build a valid descriptive model of performance. Second, the effects of individual differences can be partialled out without eliminating information as aggregation would do, either by coding subjects as a dummy variable in the regression analysis or by using the scaling criterion method suggested by Pedhazur (1982), in which each subject's mean is used as a variable in the regression (see Lee and Polson, 1989). Finally, separate regression analyses can be performed for aggregated and unaggregated data (as done by Bovair, Kieras, and Polson, 1990). Each of these techniques (and others, such as Z-transformations) allows evaluations of the power of the model both with and without the variance resulting from individual differences.

Finally, under no circumstances should one aggregate to ranges of the predictor variables in the model. This practice always eliminates variance that the model claims is directly related to performance, which reduces the amount of information available to evaluate a model. As we have demonstrated, aggregation to levels of the predictor variable inflates the value of R^2 and leads to incorrect inferences about the level of specification of the model.

Recommendations for the Use of Composite Models and Variables

Our recommendations about the use of composite variables or models are more general than those for aggregation of data. The major point is that testing a single composite model or variable with regression should not exclude the consideration of other models, no matter what the value of R^2. This is obvious. What may be less obvious is the need to provide as much information as possible about the relative contributions of the variables in the model and the ways these variables might interact to produce performance. This information is needed not only to judge the model being tested but also to build and test future models.

Information from regression analyses using composite variables or models can be maximized in several ways. One can use stepwise multiple regression and present information about F-to-enter and F-to-remove values, partial correlation coefficients, and the order in which the variables entered the equation. This has the advantage of testing multiple models, not merely the model proposed by the researchers. It also increases the amount of information available to future researchers in the area.

The use of stepwise regression is not without its drawbacks, however (see Huberty, 1989, for a review). Reports of stepwise regressions often fail to include enough information to judge the relative importance of individual variables in the analysis (e.g., testing alternative orderings for entering variables into the equation). Also, the use of a stepwise technique does not guarantee that the best possible model has been created.

Our recommendation is that when using a composite variable or a composite model in a regression, the set of correlations between each subsumed variable and the performance variables should be reported in addition to

the specific test of the model proposed by the researchers. This approach provides a maximum amount of information so that other researchers can use the analysis to confirm or disconfirm alternative models. It also avoids the drawbacks of stepwise regression. The space required in publication is more than compensated for by the additional information made available.

ACKNOWLEDGMENTS

REFERENCES

Anderson, J. R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.
Bibby, John (1977). The general linear model—A cautionary tale. In C. A. O'Muircheartaigh and C. Payne (Eds.), The analysis of survey data: Volume 2 model fitting (pp. 35–79). New York: Wiley.
Bovair, S., Kieras, D. E., and Polson, P. G. (1990). The acquisition and performance of text-editing skill: A cognitive complexity analysis. Human-Computer Interaction, 5, 1–48.
Card, S. K., English, W. K., and Burr, B. J. (1978). Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT. Ergonomics, 21, 601–613.
Card, S. K., Moran, T. P., and Newell, A. (1983). The psychology of human-computer interaction. Hillsdale, NJ: Erlbaum.
Carpenter, P. A. (1984). The influence of methodologies on psycholinguistic research: A regression to the Whorfian hypothesis. In D. E. Kieras and M. A. Just (Eds.), New methods in reading comprehension research (pp. 186–202). Hillsdale, NJ: Erlbaum.
Cramer, J. S. (1964). Efficient grouping, regression and correlation in Engel curve analysis. Journal of the American Statistical Association, 59, 233–250.
Doggan, M., and Rokkan, S. (1969). Quantitative ecological analysis in the social sciences. Cambridge, MA: MIT Press.
Egan, D. E. (1988). Individual differences in human-computer interaction. In M. Helander (Ed.), Handbook of human-computer interaction (pp. 543–567). Amsterdam: Elsevier Science Publishers.
Epps, B. W. (1986). Comparison of six cursor control devices based on Fitts' law models. In Proceedings of the Human Factors Society 30th Annual Meeting (pp. 327–331). Santa Monica, CA: Human Factors and Ergonomics Society.
Gan, K. C., and Hoffmann, E. R. (1988). Geometric conditions for ballistic and visually controlled movements. Ergonomics, 31, 829–839.
Gordon, R. A. (1968). Issues in multiple regression. American Journal of Sociology, 73, 592–616.
Haitovsky, Y. (1973). Regression estimates from grouped observations (Monograph 33). New York: Hafner Press.

Hannan, M. T. (1971). Problems of aggregation. In H. M. Blalock, Jr. (Ed.), *Causal models in the social sciences* (pp. 473–508). Chicago: Aldine.

Huberty, C. J. (1989). Problems with stepwise methods—better alternatives. In B. Thompson (Ed.), *Advances in social science methodology* (pp. 43–70). London: JAI Press.

John, B. E., and Newell, A. (1987). Predicting the time to recall computer command abbreviations. In *Proceedings of Human Factors in Computing Systems* (pp. 33–40). New York: Association for Computing Machinery.

John, B. E., Rosenbloom, P. S., and Newell, A. (1985). A theory of stimulus-response compatibility applied to human-computer interaction. In *Proceedings of Human Factors in Computing Systems* (pp. 213–219). New York: Association for Computing Machinery.

Kieras, D. E., and Polson, P. G. (1985). An approach to the formal analysis of user complexity. *International Journal of Man-Machine Studies, 22*, 365–394.

Klein, L. R. (1946). Remarks on the theory of aggregation. *Econometrica, 14*, 303–312.

Knight, G. P. (1984). A survey of some important techniques and issues in multiple regression. In D. E. Kieras and M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 112–154). Hillsdale, NJ: Erlbaum.

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence, 33*, 1–64.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data.* New York: Wiley.

Lee, A. Y., and Polson, P. G. (1989). Learning and transfer of measurement tasks. In *Proceedings of Human Factors in Computing Systems* (pp. 115–120). New York: Association for Computing Machinery.

MacKenzie, I. S. (1992). Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction, 7*, 91–139.

Mulaik, S. A. (1991). Factor analysis, information-transforming instruments, and objectivity: A reply and discussion. *British Journal for the Philosophy of Science, 42*, 87–100.

Paunonen, S. V., and Gardner, R. C. (1990). Biases resulting from the use of aggregated variables in psychology. *Psychological Bulletin, 109*, 520–523.

Pearson, E. S., and Hartley, H. O. (1958). *Biometrika tables for statisticians, volume 1.* Cambridge, MA: Cambridge University Press.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research.* New York: Holt, Rinehart & Winston.

Perloff, J. M., and Persons, J. B. (1988). Biases resulting from the use of indexes: An application to attributional style and depression. *Psychological Bulletin, 103*, 95–104.

Prais, S. J., and Aitchison, J. (1954). The grouping of observations in regression analysis. *Review of the International Statistical Institute, 22*, 1–17.

Robinson, W. S. (1950). Ecological correlation and the behavior of individuals. *American Sociological Review, 15*, 351–357.

Rushton, J. P., Brainerd, C. J., and Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94*, 18–38.

Sternberg, R. J. (1986). Time capsules are not a panacea: A reply to Vernon. *Intelligence, 10*, 277–279.

Thorndike, E. L. (1939). On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them. *American Journal of Psychology, 52*, 122–124.

Walker, N., Meyer, D. E., and Smelcer, J. B. (1993). Spatial and temporal characteristics of rapid cursor-positioning movements with electromechanical mice in human-computer interaction. *Human Factors, 35*, 431–458.

Walker, N., and Smelcer, J. B. (1990). A comparison of selection time from walking and pull-down menus. In *Proceedings of Human Factors in Computing Systems* (pp. 221–226). New York: Association for Computing Machinery.