5

Methodological Problems and Pitfalls in the Study of Human Metacognition

Bennett L. Schwartz and Janet Metcalfe

In this chapter, we discuss several methodological issues concerning metacognitive accuracy. These issues are of importance because they could serve to qualify interpretation of experimental findings concerning the nature of and mechanisms underlying people's metacognitive abilities. First, we show how the nature of the final test itself, and, in particular, the number of alternatives in that test, influences assessed accuracy of prediction. A review of the literature is given showing that accuracy of metacognitive prediction increases along with the number of alternatives at time of second test. The reasons for this strong relation are discussed. Second, we discuss how restricted range on either the judgments themselves or on the criterion variable can influence the accuracy of metacognitive predictions. We present an experiment that illustrates the impact of this potential confound. Third, we discuss problems that may arise when comparing groups that show a different mean level of problem solving, recall, or recognition. Dissociations in metacognition, occurring among such groups, may have implications for our understanding of the architecture of cognition. But we need to be confident that the dissociations are real and not a mere consequence of the methods of measurement. The use of nonparametric as compared to parametric statistics for measuring accuracy when the level of memory performance varies radically between patient groups is discussed, as are other methods of control that have been used by various researchers. Finally, following Glenberg and his colleagues we stress the importance of informing the subject as to the nature of the upcoming

memory test prior to making judgments. Given that there is no general pervasive metacognitive knowledge that is equally useful on all tests, then subjects' lack of appreciation of the nature of the test they are about to take can lead to serious distortions in assessment of their metacognitions. Subjects may know what they know, but unless they also know what kinds of questions they will be asked they may not be able to assess how well that knowledge will serve them. This lack of knowledge about the nature of the test may occur because the test conditions are not specified, because they are incorrectly specified, or because the skill level of the subject is not adequate to understanding the task.

One would like to be able to interpret differences in the accuracy of metacognitive judgments in terms of differential cognitive processes, differences in the transparency of certain tasks to conscious inspection, or differences, perhaps, among different patient groups, in mechanisms by which judgments are made. However, such cognitive interpretation may not always be straightforward because some methodological issues may cloud the results. In the studies that we discuss here subjects make one or more predictive judgments about their own future performance on a test of memory, problem solving, or comprehension. They are then given the final test and the relation between the predictive judgments and the final test results are the metacognitive "accuracy" measure of interest. We will deal mainly with experiments that investigate micropredictive accuracy or discrimination, that is, does the subject know on which questions he or she will do well or poorly? This micropredictive accuracy is often measured by a correlation coefficient, such as the nonparametric Goodman-Kruskal gamma measure, that relates the ranking of the questions to the correctness of response on the final test. In some instances, though, we will refer to studies in which macroprediction, or calibration, is of interest. In this case the prediction is one of how good performance will be, overall, without respect to which particular questions contribute in which way to that performance. Micro- and macroprediction are both metacognitive indices, but measure aspects that may be but are not necessarily linked (see Yates, 1990, p. 57). In this first section we discuss several factors concerning the relation between the judgments and the subsequent test that, failing careful consideration and control, could produce spurious results.

Number of Test

A simple factor
feeling-of-knowi
with the mechan
ber of alternativ
does not know t
he or she does
FOK rating. If h
test, this corresp
performance wil
an appropriate
the chance that
on the number
test is a two-alter
correct answer w
alternatives beco
swer become in
responses should
metacognitive ac
are expected wit
lower gamma co

Leonesio and
structed a 19-alt
number of alter
occurs in a forc
number of distra
alone. In turn, th
accuracy by redu
mance on items
rationale seems
number-alternati
idea was not, th
Leonesio, Landw
to an eight-alter
same task. Nume
alternative task (
the researchers w

Number of Test Alternatives

A simple factor that can influence the micropredictive accuracy of feeling-of-knowing (FOK) judgments, for reasons having little to do with the mechanisms underlying human metacognition, is the number of alternatives presented in the final test. Suppose the subject does not know the answer to a particular question and knows that he or she does not know, and hence gives the question a very low FOK rating. If he or she then gets that question wrong on the final test, this correspondence between the low rating and the incorrect performance will contribute to a positive correlation. This would be an appropriate demonstration of accurate metacognition. However, the chance that the person will get such a question wrong depends on the number of alternatives presented at the time of test. If the test is a two-alternative recognition test, then the odds of picking a correct answer with no knowledge are quite high. As the number of alternatives becomes large, the chances of guessing the correct answer become increasingly small. Insofar as these correct guessing responses should decrease the correlation, which is the measure of metacognitive accuracy, and insofar as more of these correct guesses are expected with fewer test alternatives, we would expect to find lower gamma correlations with few than with many alternatives.

Leonesio and Nelson (1990) noted this possibility, and hence constructed a 19-alternative forced choice test. They say: "The large number of alternatives helped to reduce the noise that necessarily occurs in a forced-choice recognition test; that is, the greater the number of distractors, the less likely an item is to be correct by chance alone. In turn, this reduces the noise in the measures of metamemory accuracy by reducing the likelihood of chance recognition performance on items that the person does not know" (pp. 465–466). The rationale seems reasonable, but the experiment used only a fixed-number-alternative forced-choice test so the empirical validity of this idea was not, thereby, demonstrated. In an earlier study, Nelson, Leonesio, Landwehr, and Narens (1986) compared a four-alternative to an eight-alternative forced-choice test on what was otherwise the same task. Numerically, the gamma values were higher on the eight-alternative task (.33) than on the four-alternative task (.28). However, the researchers were interested in other effects, and the significance