Neuropsychology
2002, Vol. 16, No. 3, 434–439

0894-4105/02/$5.00 DOI: 10.1037//0894-4105.16.3.434

Deficiencies in the Detection of Cognitive Deficits

Julian R. Keith and Antonio E. Puente
University of North Carolina at Wilmington

Numerous studies purport to show that cardiopulmonary bypass (CPB) surgery is associated with persistent postoperative cognitive decline. In J. R. Keith et al. (2002), the authors argued that reports of post-CPB cognitive declines have often been quantified using data analysis methods that were based on tenuous assumptions and overlooked problems associated with familywise Type I errors. Four peers who are recognized for their expertise in neuropsychological outcomes research evaluated the arguments developed in the J. R. Keith et al. article, critiqued the study presented in that article, and offered suggestions for how to investigate whether cognitive decline occurs reliably after CPB. In this reply article, the authors respond to the open-peer commentaries made regarding the J. R. Keith et al. study.

We are pleased that Chelune (2002), Sawrie (2002), Smith (2002), and Millis (2002) agreed to comment on our article (Keith et al., 2002). Their commentaries each addressed issues that apply specifically to the analysis of cardiopulmonary bypass (CPB) surgery effects on cognitive performance and to the research on neuropsychological outcomes in general. In our reply to the open-peer commentaries, we first focus on some of the themes that the commentaries had in common and then present a reappraisal of the study featured in the Keith et al. article.

Clearly, we are in agreement with Chelune (2002), Sawrie (2002), Smith (2002), and Millis (2002) with respect to the most fundamental issues. One point on which we agree with them is that neuropsychologists must play a key role in informing health care professionals and patients about whether certain diseases, such as coronary artery disease, and surgical interventions, such as CPB, are associated with cognitive impairment. For better or worse, clinical neuropsychology has focused on differential diagnoses of cognitive impairment after frank neurological insults (e.g., head injury or stroke). Without question, however, neuropsychology has the potential to contribute substantially to the analysis of a broad range of factors that alter cognitive functioning. Another view that we share with these authors is that a correct understanding of whether CPB is reliably associated with cognitive decline hinges on the validity of the assumptions underlying researchers' analyses of the data. Additionally, we seem to agree with them that critically important issues that are involved in the analysis of the effects of CPB on cognitive performance have been overlooked in previous studies on this topic.

Julian R. Keith and Antonio E. Puente, Department of Psychology, University of North Carolina at Wilmington.

Correspondence concerning this article should be addressed to Julian R. Keith, Department of Psychology, University of North Carolina at Wilmington, 601 South College Road, Wilmington, North Carolina 28403–5612. E-mail: keithj@uncwil.edu

Working in the Noise: Investigating Subtle and Rare Neuropsychological Phenomena

Many of the phenomena that neuropsychologists investigate are subtle, rare, and multiply determined. In such cases, the challenge lies in distinguishing genuine phenomena from ambient (random) noise or nuisance factors. Deciding whether CPB is associated with cognitive decline could serve as the prototypical example of this situation. Indeed, the pre- to postoperative changes in cognitive performance that are observed in the majority of CPB patients fall well within the range of the change scores that are produced by the non-CPB control participants. However, the hypothesis that CPB is associated with cognitive decline should not be dismissed. A fundamental question that motivated us during the Keith et al. (2002) study was how strong is the evidence against the null hypothesis? Or, to frame the question differently, how far out on the limb were we going by claiming that CPB causes persistent declines in cognitive performance? Indeed, the *p* values associated with the inferential statistics that we computed in the Keith et al. study provided just that information.

In our view, the Keith et al. (2002) study adds to the literature on the topic of CPB effects on cognitive performance, first and foremost, because the analyses demonstrated that new cognitive performance differences between the CPB and the healthy older control group appeared postoperatively on two measures of attention, and the reliabilities of these differences were confirmed using conventional inferential statistics. Furthermore, the new postoperative group differences could not be accounted for in terms of general psychomotor slowing or a loss of fine motor dexterity because performances on simple reaction time and rotor pursuit tasks were unaffected by CPB. So, the postoperative group differences on two measures of attention were likely due to genuine differences between the two groups at the cognitive level rather than at the noncognitive, sensorimotor level.

Chelune (2002), Sawrie (2002), and Smith (2002) all agreed that the analyses we used in the Keith et al. (2002) study were the appropriate ones for detecting differences at

the group level. They also noted, however, that nothing prohibits researchers from specifying whether a particular individual declined significantly as a result of undergoing CPB. We agree but also think that establishing whether the basic phenomenon occurs reliably at the group level should be a high priority. We reasoned that if cognitive decline truly is more probable in the CPB group than it is in the control group, then the reliability of this difference should be readily demonstrated using group level inferential statistics. Certainly, there is no reason why analyses that were designed to determine how many individuals within each group actually meet an investigator's criterion of being defined as *impaired* could not also be used to provide further details on the pattern of results obtained. As it has been applied to the analysis of postoperative cognitive impairment in CPB patients, however, incidence reports have been fraught with difficulties.

The Multiplicity Problem

Procedures for classifying individual participants as impaired versus not impaired, such as the standard deviation (SD) and reliable change index (RCI) methods, have been presented as if they are alternatives to traditional null hypothesis statistical testing (Kneebone, Andrew, Baker, & Knight, 1998; Stump, James, & Murkin, 2000). On the contrary, such methods test the null hypothesis for each participant individually on each measure of cognitive performance and therefore are the most extreme possible example of null hypothesis testing. As such, researchers who use methods such as the RCI should be particularly concerned about the pitfalls involved in null hypothesis statistical testing.

We were pleased that the commentaries of Chelune (2002), Sawrie (2002), and Smith (2002) all strongly reinforced our view that controlling for familywise Type I error inflation, an issue that is sometimes referred to as the *multiplicity problem*, is essential when methods such as the RCI are used to generate incidence rates. Smith noted that in Keith et al. (2002) we used "seven measures in this study, so by chance alone, 31% of the sample would be expected to show at least one measure with real decline as defined by the RCI" (p. 432). It could be argued that, in truth, Smith underestimated the number of individuals that should be expected to decline significantly on at least one measure because his estimate did not take into consideration the fact that the null hypothesis was tested individually on each participant. Thus, in using methods such as the RCI, the null hypothesis is tested $n \times m$ times, such that n is the study sample size and m is the number of cognitive measurements taken on each participant. To our knowledge, the fact that the SD and RCI methods are forms of null hypothesis statistical testing has not been previously acknowledged, much less reckoned with, in the literature on CPB effects on cognitive performance. Regardless of whether a study involves testing the null hypothesis at the group or individual case level, it is critical to recognize the need to control for familywise Type I error inflation. Toward this end, as Sawrie noted, it is also helpful to reduce the number of cognitive measures to the lowest possible number.

Practice Effects

The tendency for participants' performances to improve as a function of repeated cognitive assessment complicates interpretations of results in studies that involve serial assessment (McCaffrey, Duff, & Westervelt, 2000). One issue that has been ignored in the literature on CPB and cognitive performance is that the practice effect problem is compounded when the study uses serial assessments to compare performance changes over time in participants who are drawn from different populations. We consider it risky to assume, as the RCI method does, that practice effect sizes should be equivalent across CPB and non-CPB groups. Sawrie (2002) argued, convincingly in our opinion, that the assumption that even individuals from the same group may experience the average practice effect is untenable. Unlike the RCI, the standardized regression-based (SRB) method recommended by Sawrie predicts each participant's practice effects on the basis of his or her own preoperative score and a set of covariates chosen by the investigator, such as age, education, and gender.

Sawrie's (2002) discussion of the advantages of the SRB method over the RCI was compelling. The fact remains, however, our understanding of the determinants of individual differences in practice effect sizes is quite limited. In the case of CPB patients, they may show smaller practice effects than healthy control participants show because of preexisting learning impairments, as suggested by our finding of preoperative verbal learning impairments in CPB patients in the Keith et al. (2002) study. Other possibilities, however, should be considered as well. For example, it is well established that recently learned information is quite vulnerable to disruption by factors that interfere with normal cerebral functioning, presumably because long-term memory formation depends on memory consolidation processes that take place over the course of many days after information is initially acquired. Many examples of factors that produce retrograde amnesia for recently learned information are familiar to neuropsychologists, including electroconvulsive shock, exposures to certain drugs, and closed-head injuries. It would not be unreasonable to hypothesize that factors involved in CPB (i.e., deep anesthesia; hypothermia; exposure to large doses of cholinergic, opiate, and benzodiazepine drugs; and cerebral ischemia) may disrupt memory consolidation processes that are necessary for long-term retention of information that was learned during a preoperative cognitive assessment session that occurred 24–48 hr prior to surgery. By this account, smaller practice effects of the CPB participants may not be due exclusively to demographic characteristics of the group, preoperative learning impairments, or postoperative cognitive impairments. Rather, the CPB procedure may produce retrograde amnesia for information acquired during the days leading up to surgery.

Obviously, what is lacking here is a detailed understanding of the causes of individual, and group, differences in

practice effect sizes, and few of the plausible hypotheses have been eliminated on empirical grounds. Toward this end, Smith (2002) proposed carrying out multiple preoperative sessions to establish a stable baseline of performances in each participant prior to surgery and then carrying out multiple postoperative assessments of learning to determine whether learning impairments in CPB participants are enduring. Unfortunately, because CPB is often carried out within 1 or 2 days of diagnosis, there is not enough time to carry out multiple preoperative sessions.

Neuropsychology Research and Evidence-Based Medicine: Where Should Researchers Draw the Line?

In the Keith et al. (2002) article, we critiqued the two methods most widely used in CPB studies for defining individual patients' postoperative cognitive performances as impaired versus unimpaired, the SD and RCI methods. The central point we intended to make was that neither of these methods, as they had been applied in the literature to date, permit researchers to decide whether CPB is reliably associated with cognitive change. On the basis of our understanding of Chelune's (2002) comments, it seems to us that Chelune agreed that previous studies of CPB effects on cognitive performance that used SD and RCI methods likely overreported the incidence of postoperative cognitive impairment because they failed to control for familywise Type I errors, but this problem could be easily remedied in future studies using the appropriate mathematical corrections. Furthermore, Chelune argued that methods such as the SRB norms method that we discussed earlier (see also Sawrie, 2002) would be superior to the SD and RCI methods at identifying reliable cognitive change at the level of individual participants and could be applied in the context of CPB research. Finally, Chelune suggested that because the science of neuropsychology is dominated by the health care system, research results that are not packaged in a consumer friendly form (i.e., incidence reports) are of little value to clinicians. We think that these issues warrant further discussion.

Central to Chelune's (2002) concern, as we understand it, is the question of whether individual participants' cognitive change scores should be treated either as continuous variables or as discrete outcomes (i.e., impaired versus not impaired). Although the goal of our study was to establish whether cognitive differences between CPB patients and healthy older controls were indeed reliable, clinicians could benefit from information specifying the percentage of the study sample in whom new postoperative cognitive impairments were detected. Chelune pointed out that there are numerous quantitative methods that researchers could use to do this. However, we are hesitant to adopt these methods for two reasons.

First, before we define clinical meaningfulness in statistical terms, such as reliability and magnitude of a change score on a particular cognitive test, it must be demonstrated that these parameters are related to clinical outcomes. As we noted in the Keith et al. (2002) article, small performance changes in one cognitive domain may produce larger functional impairments in everyday life than larger changes in other cognitive domains may produce. Indeed, in our view, research specifically focused on the quantitative relationships between measures of cognitive performance and adaptive functioning in everyday life would provide information that would greatly enhance the application of research results to clinical practice. Second, by treating cognitive change scores as discrete outcomes, as the SD, RCI, and SRB methods do, researchers are discarding potentially valuable information regarding the effects of CPB on cognitive functioning that did not reach an arbitrary threshold for being classified as impairment. By treating cognitive performance (and, by extension, impairment) as a continuous variable, our analyses in the Keith et al. study counted all degrees of change (positive and negative) as meaningful. We make no pretense of knowing which particular cognitive performance changes should be treated as most important.

The vast majority of studies on the effects of CPB on cognitive performance have provided reports of how many patients showed evidence of postoperative decline, albeit using flawed quantitative methods, without establishing the general reliability to the phenomenon under investigation. In the Keith et al. (2002) study, it seemed to us that priority should be given to establishing whether cognitive performance differences between CPB and healthy control participants were actually reliable. Additionally, it seems clear that until researchers have a better understanding of how performance on cognitive tests relates to participants' performances outside of the laboratory (i.e., their ecological validity) incidence reports may be limited in terms of how much information they actually convey regarding the clinical meaningfulness of CPB-related cognitive changes.

As a compromise between presenting a count of the number of patients that meet a fixed criterion for being defined as impaired and presenting inferential statistics that omit detailed information about the patterns of change observed consider the data presented in Figure 1. Figure 1 presents our data from the Keith et al. (2002) study in a novel way. Each panel in Figure 1 represents the distribution of *z*-score changes (postoperative minus preoperative *z* score) in controls (left side, light gray bars) and CPB patients (right side, dark gray bars) on each individual task. Pre- and postoperative *z* scores were computed separately as described in the Keith et al. article, thus removing the influence of practice effects (keeping in mind the caveats involved in our assumptions about practice effects that were discussed earlier). Scores above zero represent postoperative improvement (relative to the entire study sample distribution) and scores below zero represent postoperative decline. The dark horizontal lines that extend out from the vertical axes in each panel indicate the 99% confidence interval (CI) for the control group's means. Thus, it could be argued, using the same reasoning that is applied when RCI and SRB methods are used, the regions of the data distributions that fall below the lower 99% CI indicate the percentage of participants whose postoperative change scores were significantly lower than the mean of the control group's mean change score. An important advantage of

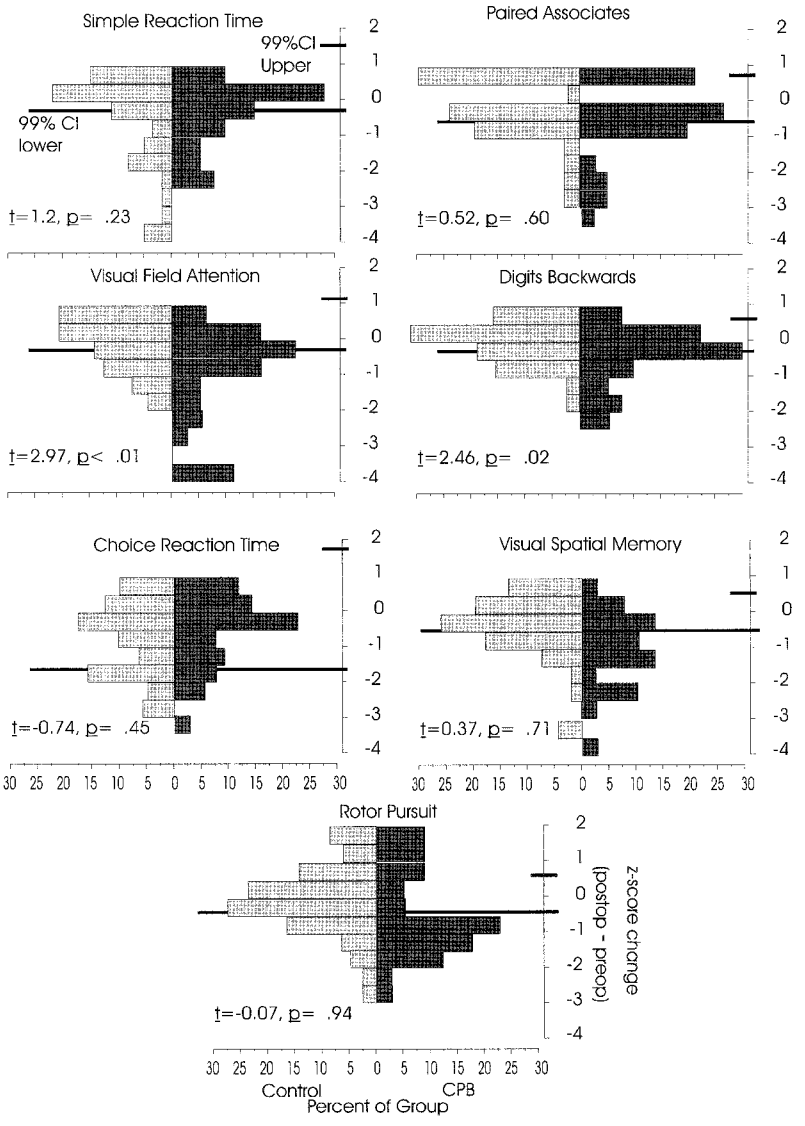

Figure 1. Each panel shows data from a single cognitive test and organizes the distributions of *z*-score changes (postoperative *z* score minus preoperative *z* score; shown on the *y*-axes) in bins that span 0.5 standard deviations. The *x*-axes in each panel indicate the percentage of each group (cardiopulmonary bypass [CPB] group appears in dark gray on the right, and control group appears in light gray on the left) represented by each bar in the figure. CI = confidence interval; postop = postoperative; preop = preoperative.

organizing the data in this manner is that the data show the distributions of postoperative changes in the CPB group's performances on each cognitive task relative to the pattern of change observed in the control group. Important to note on each task, as can be seen in Figure 1, a substantial percentage of both groups had change scores that fell outside the 99% CI. Thus, the distribution of change scores observed in the control group provides an index of the expected false-alarm rate. Finally, when the data are presented as shown in Figure 1, researchers can decide for themselves where they think that the threshold for impairment should be placed and guide their decision using the data distribution produced by the control group based on the false-alarm rate that they are willing to tolerate.

A Reappraisal of the Keith et al. (2002) Study

> Worship the spirit of criticism. If reduced to itself, it is not an awakener of ideas or a stimulant to great things, but, without it, everything is fallible; it always has the last word.
> —Louis Pasteur (as quoted in Vallery-Radot, 1923)

The Keith et al. (2002) study was designed to determine whether (a) cognitive performance was adversely affected in patients who underwent CPB, (b) cognitive impairment

was present in the surgery candidates prior to CPB, and (c) cognitive impairments in CPB patients were domain specific. In addition to evaluating the arguments presented in the Keith et al. (2002) article regarding quantitative methods used to analyze CPB effects on cognitive performance, Chelune (2002), Sawrie (2002), Smith (2002), and Millis (2002) all critiqued the design and analyses we used in the study. We welcome the opportunity to reply to some of their criticisms of our study.

Subject attrition has been a common problem in studies of the effects of CPB on cognitive performance. In this respect, the Keith et al. (2002) study is no exception, with 17 CPB and 6 control participants dropping out of the study. Additionally, 1 CPB participant's data were excluded from the analyses because his postoperative declines were at least four SDs larger than the mean of the CPB group on all measures, and thus, his data were clearly not representative of the overall pattern observed in that group. The commentaries by Smith (2002) and Millis (2002) pointed to many of the reasons why subject attrition occurs in studies of this sort, and both authors suggested that the analysis of our results would have been enhanced if the missing data had been replaced with estimates generated using data imputing methods. Millis briefly mentioned some of the different algorithms that have been developed for imputing missing data. Critical here, however, is the fact that a researcher's choice of a particular data imputation method depends on whether he or she can determine whether (a) the data is missing completely at random, (b) the missing data is predictable from other variables in the database, or (c) the missing data is not random and not predictable from other variables in the database. The full information maximum likelihood (FIML) and multiple imputation (MI) methods that were recommended by Millis assume that the data are missing completely at random (Little & Rubin, 1987). As Millis noted in his commentary, however, missing data in studies of CPB patients may be due to any number of factors that are not random. Thus, the primary assumption of the FIML and MI methods would not be met under the present circumstances. To be sure, other methods are available for imputing missing data that do not assume complete randomness (Schafer, 1997). Ultimately however, as Smith stated, although modeling the different possible patterns of outcomes under various assumptions may be an interesting exercise, it is impossible to validly know the effects of subject attrition.

Chelune (2002), Sawrie (2002), Smith (2002), and Millis (2002) all took issue with the use of healthy older participants as controls and argued that medically managed coronary vascular disease patients (Chelune, 2002, p. 423), other surgery groups (Millis, 2002, p. 426; Smith, 2002, p. 433), and CPB candidates randomly selected to not undergo CPB (Sawrie, 2002, p. 430) should have been studied. Undoubtedly, more information about the causes of postoperative cognitive changes in CPB patients would be gained by including control groups that were matched on various dimensions. The study design that we used in the Keith et al. (2002) study would not permit researchers to conclude that postoperative cognitive performance changes in CPB patients were caused by factors unique to CPB surgery. That is why we cautioned researchers against interpreting the data in such a manner (Keith et al., 2002, p. 418). Factors such as the presence of coronary artery disease, receipt of general anesthesia, invasion of the thoracic cavity, and pre- and postoperative psychological stress, to name a few, all may play roles in producing the cognitive performance differences observed between CPB and healthy controls.

Analytically, the most powerful way to determine whether factors unique to CPB cause cognitive decline would be to do a true experiment and randomly assign participants to groups that receive either CPB or sham CPB surgery (i.e., anesthesia, sternotomy, hypothermia, etc.; the most powerful design would include experimental groups that were exposed to different subsets of these factors). Although imagining the ideal experiment is a useful intellectual exercise because it helps researchers to conceptualize the various possible factors that may influence post-CPB cognitive changes, in reality, such an experiment would not be ethical because it would create unwarranted risks for participants.

Quasi experiments, in which comparison groups are defined in terms of a factor that is not controlled by the experimenter (i.e., surgery group, age, gender, etc.), provide another method for investigating how different aspects of the CPB procedure affect cognitive performance. To paraphrase Sawrie (2002), an analysis is only as good as the control group that it is based on. Ideally, therefore, the control group and the CPB group should be equivalent in every respect except along the dimension of interest (i.e., surgery). Researchers might argue, therefore, as Sawrie, Chelune (2002), and Millis (2002) did, that healthy older participants are a poor choice as a control group because they differ from CPB patients in at least two respects, their preoperative health status and CPB surgery. Furthermore, Chelune stated

> More careful consideration of the assumptions underlying the choice of an appropriate control group might have led Keith et al. to choose a cardiac control group who either elected to undergo pharmacologic management of their condition or were simply wait listed and tested twice before undergoing surgery as has been done with other patient groups. (p. 423)

At first glance, Chelune's (2002) suggestions seem straightforward. In practice, however, we encountered several unanticipated complications when we pursued such strategies during the Keith et al. (2002) study. Over a 4-year period, we solicited coronary artery disease patients scheduled for coronary artery angioplasty and participants undergoing non-CPB surgeries. Our reasoning was like that of the commentators; we assumed that such groups would provide appropriate controls for effects of coronary artery disease and of general surgery procedures (e.g., anesthesia) on cognitive performance. However, relative to our experiences recruiting healthy control participants and CPB patients, few individuals in the angioplasty or general surgery conditions agreed to participate in the study. Furthermore, among those who did enroll, attrition rates were more than twice that of the CPB group. Most important, however,

baseline cognitive performances of angioplasty and general surgery patients were not as well matched with the performances of the CPB group as were those of the healthy older controls. In a study on the effects of CPB on cognitive performance, it would seem that the most important dimension on which to match the groups would be preoperative cognitive performances. In the end, it is important to note that on six of seven measures of cognitive performance used, the CPB patients and the healthy older control participants were very closely matched during preoperative baseline testing, permitting a meaningful analysis of postoperative differences between the groups.

As mentioned earlier, one of our goals in the Keith et al. (2002) study was to determine whether certain cognitive processes were affected more than others were by CPB. On the basis of the findings that controls outperformed CPB patients postoperatively on two very different sorts of measures of attention, visual spatial attention (a reaction time based task) and backwards digit span (Wechsler, 1981), but not on five other performance measures, we concluded that attention may be particularly vulnerable to CPB effects. Millis (2002) challenged this interpretation of our data, arguing instead that the tasks used to measure cognitive performances may have differed in terms of the tasks' sensitivities to cognitive change. Millis's point is well taken.

Millis (2002) highlighted some of the difficulties involved in assessing domain specific cognitive performance changes that should concern neuropsychologists. In fact, as we considered the implications of Millis's argument, it occurred to us that the matter is considerably more complicated than it seems on first impression. For example, Millis made the valid point that instruments that are not equivalent in terms of test–retest reliability then are not equivalent in terms of their sensitivities to cognitive change. In the Keith et al. (2002) study, however, in which our aim was to compare the effects of a factor (CPB) on cognitive processes in different domains, the problem is not only whether tests used to measure cognitive performances in the different domains produce different levels of measurement error, but also whether the particular cognitive processes that these tests measure differ from one another in terms of the levels of variability intrinsic to each. That is, we do not know whether the cognitive processes that support attention are more or less stable (i.e., reliable) than those that support memory, perception, problem solving, and so forth. Furthermore, we wonder whether it is possible, even in principle, to dissociate genuine variance in cognitive processing from variance due to measurement error. These issues present interesting methodological challenges to those involved in investigating domain specific cognitive decline.

Finally, we would like to close by acknowledging that the commentaries by Chelune (2002), Sawrie (2002), Smith (2002), and Millis (2002) were immensely valuable to us because they stimulated us to reevaluate many of our own assumptions and practices regarding the study of cognitive change over time and to reconsider the implications of our data presented in the Keith et al. (2002) study. We hope that the readers of *Neuropsychology* benefit as much as we have from this frank and open exchange of ideas.

References

Chelune, G. J. (2002). Making neuropsychological outcomes research consumer friendly: A commentary on Keith et al. (2002). *Neuropsychology, 16,* 422–425.

Keith, J. R., Puente, A. E., Malcolmson, K. L., Tartt, S., Coleman, A. E., & Marks, H. F., Jr. (2002). Assessing postoperative cognitive change after cardiopulmonary bypass surgery. *Neuropsychology, 16,* 411–421.

Kneebone, A. C., Andrew, M. J., Baker, R. A., & Knight, J. L. (1998). Neuropsychologic changes after coronary artery bypass grafting: Use of reliable chance indices. *Annals of Thoracic Surgery, 65,* 1320−1325.

Little, R. J. A., & Rubin, D. A. (1987). *Statistical analysis with missing data*. New York: Wiley.

McCaffrey, R. J., Duff, K., & Westervelt, H. J. (2000). *Practitioner's guide to evaluating change with neuropsychological instruments*. New York: Kluwer Academic/Plenum Publishers.

Millis, S. R. (2002). Measuring change: A commentary on Keith et al. (2002). *Neuropsychology, 16,* 426–428.

Sawrie, S. M. (2002). Analysis of cognitive change: A commentary on Keith et al. (2002). *Neuropsychology, 16,* 429–431.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* New York: Chapman & Hall.

Smith, G. (2002). What is the outcome we seek? A commentary on Keith et al. (2002). *Neuropsychology, 16,* 432–433.

Stump, D. A., James, R. L., & Murkin, J. M. (2000). Is that outcome different or not? The effect of experimental design and statistics on neurobehavioral outcome studies. *Annals of Thoracic Surgery, 70,* 1782−1785.

Vallery-Radot, R. (1923). *The life of Pasteur*. Garden City, NY: Doubleday.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised Manual.* San Antonio, TX: Psychological Corporation.

Received March 4, 2002
Accepted March 4, 2002 ■