Predicting Memory
for Childhood Sexual Abuse:
"Non-Significant" Findings with the Potential
for Significant Harm

Eileen L. Zurbriggen
Kathryn Becker-Blease

In the March 2003 issue of *Psychological Science*, the flagship journal of the American Psychological Society, Goodman et al. (2003) reported on a prospective study that examined memory for childhood sexual abuse. The authors interviewed adolescents and young adults who had been victims of documented childhood sexual abuse that led to criminal prosecution for the perpetrator. This sample was a subset of those who had previously (approximately

Eileen L. Zurbriggen, PhD, is Assistant Professor in the Department of Psychology, University of California, Santa Cruz. Her research interests include sexual aggression perpetration, adult outcomes of childhood sexual abuse victimization, and societal and intrapsychic links between power and sexuality. Kathryn Becker-Blease, PhD, Developmental Psychology, is a post-doctoral fellow at the Family Research Laboratory at the University of New Hampshire. There she is involved in research on Internet victimization of children and studies of the prevalence of child victimization in the United States. Her other research interests include the intergenerational transmission of trauma, and attention, memory and learning of abused and traumatized children.

Address correspondence to: Eileen L. Zurbriggen, Department of Psychology, University of California, Santa Cruz, Santa Cruz, CA 95064 (E-mail: zurbrigg@ucsc.edu).

13 years earlier) participated in a study of court testimony (Goodman et al., 1992). As part of the follow-up study, participants were asked to report any experiences of childhood sexual abuse. Failure to report the target incident could indicate lack of memory for the abuse.

Very few prospective data sets concerning memory for abuse are available; therefore, this is an important study. Prospective studies can potentially add to our understanding of the prevalence of impairment, distortion, and loss of memory for childhood trauma as well as increase our understanding of the mechanisms whereby such distortion or loss might occur. This latter goal was pursued by Goodman et al. (2003) through the examination of correlational relationships between disclosure of abuse during the follow-up study and such variables as gender and age of the victim, relationship between perpetrator and victim, and presence of maternal support for the victim. There were two main findings. First, the non-disclosure rate was substantially lower than in two previous prospective studies (Widom & Morris, 1997; Williams, 1994). Goodman et al. concluded, therefore, that forgetting childhood abuse might be relatively uncommon. The second major finding was that age, severity of abuse, and maternal support were related to disclosure, but gender, ethnicity, relationship with perpetrator, and extent of legal involvement were not. Goodman et al. used this latter finding to discount a variety of theories related to memory for trauma and abuse.

There are many possible reasons why the non-disclosure rate might have been lower in this study. For example, one would expect better memory for abuse among survivors who had the opportunity to talk repeatedly about their victimization experience(s) and who were believed by caregivers and authority figures (see Freyd, this issue). Thus, it is important not to generalize this first finding inappropriately.

It is also important to carefully assess the analyses related to the second finding, especially given the authors' broad arguments for rejecting several theories related to memory for trauma. For example, Goodman et al. (2003) stated "These findings do not support the existence of special memory mechanisms unique to traumatic events, but instead imply that normal cognitive operations underlie long-term memory for CSA. The role of dissociation, however, deserves further study" (p. 117). The implication is that dissociation is the only "special" mechanism deserving of study. This is a strong conclusion to draw from one study, even one with definitive data. If there are concerns about the interpretation of the statistical analyses, the conclusion is even more debatable. We argue below that because of low statistical power, sparseness of data, and possible multicollinearity between predictors, Goodman et al.'s analyses should be interpreted with caution. The substantive conclusions they draw may be premature.

STATISTICAL POWER

In several places Goodman et al. (2003) highlighted the theoretical importance of some of the predictor variables (e.g., "examine several additional potential predictors of long-term memory for abuse that are of substantial theoretical importance," p. 114; "several additional variables were examined because of their theoretical importance," p. 116). Results for variables that did not reach the standard ($p = .05$) level of statistical reliability were then presented briefly and in summary form (e.g., "Legal involvement, gender, and perpetrator relationship were not significant predictors," p. 116, and "None of these variables was significantly related to disclosure," p. 116). This presentation implies that the failure to reject various null hypotheses can be used as evidence that these null hypotheses are true (e.g., to conclude that level of relationship betrayal has no impact on probability of disclosure). However, this logical conclusion is justifiable only if one can demonstrate that statistical power for the analysis is high (Cohen, 1990). Typically one would prefer power to be at least .80 (20% chance of making a Type II error) before drawing such conclusions. Concerns about power are important for any statistical analysis; however, they are of particular importance here due to the fact that Goodman et al. use a lack of significant results to reject certain theoretical perspectives. Before accepting these conclusions, it is vital to demonstrate that their study had adequate power to detect population effects that might exist. If power were low, it would not be appropriate to imply that the null hypotheses are true. It would be especially inappropriate to reject entire theoretical perspectives on the basis of a failure to reject individual null hypotheses in one study.

The logistic regression analysis conducted by Goodman et al. (2003) is likely to have low power for several reasons. First, there were a relatively large number of predictors. Maxwell (2000) has recently argued that most regression analyses are underpowered. He notes that power for testing an overall effect (i.e., does the overall R^2 differ from zero?) is greater than the power for testing the effect of an individual predictor (i.e., does the slope of an individual partial regression coefficient differ from zero?). For example, Maxwell notes that in an analysis with five predictors that are moderately ($r = .30$) correlated with each other and with the dependent variable, a test of an individual regression coefficient requires a sample size of 419 to achieve 80% power. A larger sample will be required to achieve the same power if (a) more predictors are used, (b) the correlations between predictor variables are higher than .30, or (c) the correlation between the predictor of interest and the dependent variable is lower than .30. Goodman et al. used seven predictors in their logistic regression analysis. Although they did not present a zero-order correlation matrix,

the other two conditions are also likely to be true. Thus, even more than 419 participants would be necessary to reach 80% power.

There are additional reasons to believe that this analysis is seriously under-powered. Most of the dichotomous predictor variables (i.e., age, relationship with perpetrator, maternal support, ethnicity) have proportions that differ dramatically from 50/50; Whitehead (1993) has shown such cases to result in reduced power. The overall proportion for the dependent variable also differs markedly from 50/50 (26 non-disclosers vs. 142 disclosers), which also results in diminished power (Hsieh, 1989). Power is further reduced because the data are sparse (see below).

Therefore, it would be inaccurate and misleading to interpret the fact that a predictor's p-value was greater than .05 as strong evidence that the predictor does not have an effect on disclosure. Yet Goodman et al. (2003) imply that the coefficients can be interpreted in this way. Even worse, they seem to be arguing that entire theoretical perspectives (e.g., betrayal trauma theory [Freyd, 1996], Terr's [1991] typological theory of trauma, various psychoanalytic theories and, indeed, all "special" memory mechanisms) have been disconfirmed, based on single non-significant coefficients in an underpowered regression analysis.

Note that two predictors came close to meeting the standard ($p = .05$) level for rejecting the null hypothesis–ethnicity ($p = .06$) and legal involvement ($p = .13$). Goodman et al. (2003) provided some discussion about ethnicity, but not about legal involvement. This is curious because although there are well-developed theoretical accounts for why greater legal involvement might lead to improved memory (e.g., elaboration, construction of schemas and scripts, simple rehearsal), a theoretical account of the effect of ethnicity is more tentative. Goodman et al. stated that the "most parsimonious" explanation is a "lack of willingness to disclose" (p. 117) by African Americans. It is not clear why this is a parsimonious explanation (more parsimonious than what?) or even the most plausible explanation (because no motivation for the lack of willingness is given). An equally plausible explanation would directly implicate racism, especially as present in the legal justice system. If prosecutors and others expressed more doubts about the veracity of the accounts given by African American children and if they provided less support to the families of these children, this may have led to impaired memory for the target incident.

Perhaps the reason that Goodman et al. (2003) did not discuss legal involvement is that in the logistic regression the p value for this variable was greater than .10. The authors did not report analyses of the zero-order relationship between legal involvement and non-disclosure; however, from the data they report it is possible to compute cell sizes (disclose vs. non-disclose) for each of the three categories of legal involvement (testify: 47 vs. 4; courthouse: 49 vs.

10; no appearance: 46 vs. 12). If one conducts a correlational analysis with these data, the results are in the theoretically predicted direction. Greater legal involvement (more talking, more rehearsing) is related to disclosure, and the effect nearly reaches the standard level of statistical significance (point biserial $r = .14$, $n = 168$, $p = .07$).

Another theoretical perspective that was rejected by Goodman et al. (2003) is betrayal trauma theory (Freyd, 1996). The justifications for this rejection apparently included the results of the logistic regression analysis as well as the results of a zero-order correlational analysis. Goodman et al. did not report whether the correlation between perpetrator relationship and disclosure was positive or negative. Yet if the correlation between these two variables was in the direction predicted by betrayal trauma theory (more betrayal is related to less likelihood of disclosing), it may have been more or less the same size as the correlation found by Williams (1994). Although it is not clear whether Goodman et al. dichotomized the perpetrator relationship variable for the correlational analysis (as they did for the logistic regression analysis), doing so would very likely have reduced the size of the correlation with other variables (MacCallum, Zhang, Preacher, & Rucker, 2002). In the best-case scenario (dichotomizing at the midpoint of a normally distributed variable), the size of a correlation with a continuous DV would have been reduced to about 80% of its original value (MacCallum et al., 2002, p. 24). Williams (1994) used a four-level continuous variable to measure perpetrator relationship and found a correlation of .22. If there was a correlation in the Goodman et al. data on the order of .20 between relationship with perpetrator (measured as a continuous variable) and disclosure, we would expect this correlation to at least drop to .16 (and probably even lower) if the relationship variable was dichotomized. For a sample of this size ($n = 168$), the attenuated correlation would likely not reach the standard level of statistical significance). In that case, interpreting a null finding as a repudiation of betrayal trauma theory would be a misinterpretation of the data.

As a further consideration of the relationship betrayal factor, we asked the question: "What would Goodman et al.'s (2003) findings look like, if their sample had the exact 2 X 2 frequency pattern (for disclosing as a function of perpetrator relationship) as seen in the Williams (1994) data?" To answer this question, we used a breakdown of the Williams data that indicates the exact n's for each of these four cells (reported in Freyd, 1996, p. 149). These frequencies were (a) relative/no disclose: 19; (b) not a relative/no disclose: 25; (c) relative/disclose: 21; and (d) not a relative/disclose: 55. Thus, of those women who did not disclose the abuse, 43.2% were abused by a relative and 56.8% were abused by someone who was not a relative; of the women who disclosed the abuse, 27.6% were abused by a relative and 72.4% were abused

by someone who was not a relative. We applied these percentages to the number of non-disclosers ($n = 26$) and disclosers ($n = 142$) reported by Goodman et al. to obtain the following hypothetical frequencies: (a) caregiver/no disclose: 11; (b) not a caregiver/no disclose: 15; (c) caregiver/disclose: 39; and (d) not a caregiver/disclose: 103. If one runs a correlational analysis on these hypothetical data, one obtains $r(168) = .12$, $p = .13$. In other words, the relationship is in the predicted direction and the effect size is close to what Goodman et al. report (although we do not know exactly because they provide only a range of effect sizes that includes other variables; in addition, they do not indicate the direction of the effect). The point is that the percentages of disclosers by perpetrator type could be identical to what Williams found, but not significant (for a sample of this size with only 15% non-disclosers) if one dichotomizes. (Note that a dichotomous analysis similar to this one was reported as not statistically significant [$p = .14$] by Williams; however, her analysis using a four-point continuous measure was statistically significant [$p = .05$]).

Other Statistical Problems

There are a number of other problems with the logistic regression analysis. These problems make interpretation of the analysis difficult and suggest that the conclusions may be overstated.

Sparseness of Data. Data are said to be sparse if they do not "cover" the multivariate space defined by the predictors. In other words, some areas of the multivariate space (e.g., girls who were abused by a stranger before age 5 who did not receive maternal support) might have zero or only a few data points. This problem is conceptually similar to having empty cells in an ANOVA analysis. Following Cohen, Cohen, West, and Aiken (2003, p. 506) and considering only the categorical predictors, the logistic regression analysis conducted by Goodman et al. (2003) comprises at least 192 cells: 2 (disclose) X 2 (age) X 2 (gender) X 2 (perpetrator relationship) X 2 (maternal support) X 2 (ethnicity) X 3 (legal involvement). Given that the number of cells exceeds the number of participants, sparseness (the presence of cells with no or very few data points) is clearly present. Sparseness leads to bias in the Wald tests for individual predictors, as well as in the significance test for the overall model; analyses where sparseness is a problem "should not be trusted" (Cohen et al., 2003, p. 506).

Multicollinearity. Multicollinearity exacerbates the problem of sparseness. If some predictors are correlated, then a larger sample size is required in order to have an adequate number of data points in each "cell." To see this, consider a simple 2×2 design (e.g., predicting disclosure from gender and relationship

with perpetrator). Assuming that gender and relationship with perpetrator are statistically independent (so that the correlation between the two is zero), a sample of 40 would provide an average of 10 participants in each of the four cells (assuming equal proportions in the population for the two genders and the two types of perpetrators). If, however, the two predictors had high multicollinearity, a sample of 40 participants might produce 18 girls who had been abused by a parent, 18 boys who had been abused by a stranger, 2 girls who had been abused by a stranger, and 2 boys who had been abused by a parent. In this case, sparseness is a problem even though we in theory have a large enough sample to "cover" all of the cells. The problem would be even greater if the proportions for individual predictors differ dramatically from 50/50 (e.g., if only 25 percent of the sample was male, we might have zero or only one boy who had been abused by a parent). In the Goodman et al. (2003) data set, the proportions of most predictors are quite different from 50/50. In addition, multicollinearity between predictors is likely to be present. Thus the problem of sparseness in this data set is likely to be even more severe than is suggested by a simple calculation of the number of cells in the design.

Even if a large enough sample is available, multicollinearity can still be problematic. High levels of multicollinearity can make interpretation of regression coefficients difficult because the magnitude and sign of the coefficients can change dramatically depending on what other predictors are included in the model. Goodman et al. (2003) did not report multicollinearity statistics, nor did they provide information about relationships between predictors. Thus, there is no way to tell whether multicollinearity was a problem in the logistic regression analysis reported.

CONCLUSION

Although Goodman et al. (2003) concluded that forgetting is relatively infrequent among abuse survivors whose perpetrators went through a criminal investigation, there is compelling evidence that some survivors do forget abuse (Cheit, 1998; Freyd, 1996; Widom & Morris, 1997; Williams, 1994). It would be regrettable if the Goodman et al. findings were used as a means of casting doubt on all delayed or recovered memories of childhood abuse, even those cases in which the child had no opportunity to discuss the abuse or was not believed and supported. Adult survivors who are struggling with emerging memories need to be believed. Significant harm could result from a repetition of the denial and discounting that occurred in childhood.

Interpreting null results as evidence that the null hypothesis is true is not valid unless high statistical power has been demonstrated. In this case, statisti-

cal power was low; however, a nuanced discussion of the implications of low power was not presented. This could lead other researchers to conclude that a wide range of theories developed to explain forgetting are unsupported. Such a premature conclusion could have the unfortunate effect of limiting research in this important area.

Especially when an area of research has significant political, legal, and clinical implications, it is incumbent upon researchers to discuss their findings and the limitations of their studies as clearly as possible. In this case, the authors cast doubt on entire theoretical perspectives on the basis of a limited number of tests in a single sample. It will be regrettable if these conclusions are applied within the context of political debates, legal cases, or in clinical settings without a careful consideration of the limits of this (or any one) particular study.

REFERENCES

Cheit, R. E. (1998). Consider this, skeptics of recovered memory. *Ethics and Behavior*, *8*, 141-160.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd edition). Mahwah, NJ: Lawrence Erlbaum.

Freyd, J. J. (1996). *Betrayal Trauma: The Logic of Forgetting Childhood Abuse*. Cambridge, MA: Harvard University Press.

Freyd, J. J. (2003). Memory for abuse: What can we learn from a prosecution sample? *Journal of Child Sexual Abuse*, *12*(2), 97-103.

Goodman, G. S., Taub, E. P., Jones, D. P. H., England, P., Port, L. K., Rudy, L., & Prado, L. (1992). Testifying in criminal court. *Monographs of the Society for Research in Child Development*, *57* (5, Serial No. 229).

Goodman, G. S., Ghetti, S., Quas, J. A., Edelstein, R. S., Alexander, K. W., Redlich, A. D., Cordon, I. M., & Jones, D. P. H. (2003). A prospective study of memory for child sexual abuse: New findings relevant to the repressed-memory controversy. *Psychological Science*, *14*(2), 113-118.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine*, *8*, 795-802.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19-40.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods*, *5*, 434-458.

Terr, L. C. (1991). Childhood traumas: An outline and overview. *American Journal of Psychiatry*, *148*, 10-20.

Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in Medicine*, *12*, 2257-2271.

Widom, C. S., & Morris, S. (1997). Accuracy of adult recollections of childhood victimization: Part 2. Childhood sexual abuse. *Psychological Assessment*, *9*, 34-46.

Williams, L. M. (1994). Recall of childhood trauma: A prospective study of women's memories of child sexual abuse. *Journal of Consulting and Clinical Psychology*, *62*, 1167-1176.