PUBLIC HEALTH INSTITUTE

Center for Research on Adolescent Health and Development
2001 Addison Street, 2nd Floor • Berkeley, CA 94704 925-284-8118 • Fax 925-284-8107 • http://crahd.phi.org/

January 21, 2003

Norma Munroe
Education Programs Consultant
Healthy Start and After School Partnerships
California Department of Education
1430 N Street, Suite 6408
Sacramento, CA 95814

Dear Norma,

As promised, I will provide some brief comments on the statewide Teen Pregnancy Prevention Grant Program evaluation report (Making the Connection: How School-Community Partnerships Address Teenage Pregnancy Prevention, Cagampang, et al. 2002) and explain my view that the results regarding program effectiveness are inconclusive.

To review the context, I recently requested a copy of the evaluation report to respond to what I thought were misrepresentations of the TPPGP program and evaluation results in a draft National Campaign to Prevent Teen Pregnancy (NCPTP) case study of California teen pregnancy programs that I was asked to review. The evaluation conclusions had also come up several times in discussions that are part of the Wellness Foundation teen pregnancy prevention policy work I am currently involved in, but I had only seen an executive summary before this week. I understand that only the summary has been publicly released, but that the full report has been provided to the legislature. I appreciated that you arranged for me to quickly get a copy, and at the same time I learned that you had essentially the same concerns as I had had regarding the misrepresentation in the NCPTP case study, and already had communicated this to the author. Once I had received a copy of the evaluation report, I was asked by both you and Wade Brynelson for my reactions, and felt obligated to respond constructively and honestly. Because the issues involve are so complex, I did not feel comfortable providing these comments verbally, hence this written review. Because of time limitations, I will have to limit my comments to just the most important issues I see.

I think the essence of the report's conclusions is summarized in the first paragraph of the conclusions section of the executive summary on page 6, so I will focus on these in this review:

> *The cumulative evidence indicates that TPPGP contributed to the larger than expected decreases in teen births, encouraged more teens to delay sexual activity, and helped teens in targeted programs to talk more frequently with their parents about issues related to sexuality.*

Although I disagree with all three of these conclusions, I want to emphasize that I do not think there is evidence of the opposite conclusion (that the programs overall were ineffective), but rather the evaluation results are simply inconclusive. I also recognize that the evaluators had inherited an extremely difficult and challenging situation from which to extract conclusive results.

There are several primary reasons why I disagree with the conclusions of positive overall program effects. Most fundamentally, it is inappropriate to pool data across 37 vastly different sites, with different programs, populations, baseline rates, levels of fidelity, outcome measures, and consent rates for survey participation, among other things. At the initial evaluation advisory meeting back in 1997, I first pointed this out and advised that some sort of meta-analytic approach, not necessarily a formal meta-analysis, would be needed to aggregate site-specific findings across sites. I continued to make this point through my involvement in the second advisory committee in 2001, as well as to discuss potential compromise methods to at least model the cross-site differences in all primary analyses. The issue of inappropriately pooling data across highly diverse programs and sites is not an esoteric or debatable point. If one were to design a controlled study in which each site had the same program, population characteristics, baseline rates, levels of fidelity, outcome measures, and consent rates for survey participation, then this point would still be valid, although other solutions would be available such as multilevel models that incorporate both the site-specific variance and the cross-site shared variance, for example; and some might argue that in certain situations it would not be worth the added complexity. But in a situation such as the TPPGP, the site and program differences were so profound that it is impossible to justify pooling data across sites. This applies whether individual person data are pooled, as was done with the birth rate analyses, or aggregate site data are pooled, as with the survey analyses.

There is one place in the evaluation where I believe that the site units were dealt with appropriately, that is the figure on page 14 that displays changes in teen birth rates. Each site's data are represented as a separate bar on the chart, in a beautiful and powerful display that orders the bar by size, and with statewide data included as a comparison. Although no statistical tests are included, this is an effective and appropriate data display. But the bad news is that this figure shows that 21 sites had results less positive than the statewide average, and only 16 had results better than the statewide average. One could always argue that the statewide average is not an appropriate comparison, but the best case this argument could lead to would be a conclusion that this analysis was inconclusive.

Considering the other birthrate comparisons found on pages 10-14, I think that using the zip codes for applicant programs that were not funded is, in itself, a reasonably good strategy for building an approximate comparison group, and given the situation there were probably no better options. But the cross-site data pooling issue remains and casts strong doubts upon the interpretation of these results. Also, the inclusion of 12-13 year olds in the analysis makes no sense because, first, these rates are so low as to have very large expected error, and second, because we know that the vast majority of sexually experienced girls under 14 have been forced (e.g., according to the Alan Guttmacher Institute,1994, Sex and America's Teenagers, 74% had forced intercourse, and 61% had only forced intercourse) so any fluctuations over time in this age group are not likely to be linked to the TPPGP programs.

Hence the summary figure on page 13, which compares CA, TPPGP, and comparison communities, in my opinion, represents questionable comparisons for the three reasons

described above. But even if these questions were all put aside, the results are still inconclusive: TPPGP communities decreased 3 percentage points more than comparison communities, but 8 percentage points less than the statewide average.

Regarding the second part of the overall conclusions, that the programs *"encouraged more teens to delay sexual activity,"* first it is important to note that only about half of the programs provided data (page 6), that data quality were highly variable across sites and low overall, and that response rates were highly variable and often very low. Here the data pooling issue is dealt with differently than it was for birthrates, by creating "cases," i.e., aggregated data matched by school and grade at each site (which were all that were available based on the reporting forms used, i.e., individual student level data were not available to the statewide evaluation), and then pooling the cases. Of eight independent tests done regarding delay of sexual activity (middle school and high school girls and boys in cross-sectional and matched designs), only one was reported as significant – middle school boys, and this appears to be in error. Degrees of freedom based on the *number of students* rather than the *number of cases* appear to have been used in calculating the significance tests for the various cross-sectional comparisons throughout the evaluation (e.g., there is no other conceivable way that such extreme p-values, such as .0007 for middle school boys, could have been obtained as those reported given the size of the differences). This choice is fundamentally incorrect. Yet even if this were not an error, the finding of one significant difference across the eight age/gender/data-type groups tested would not support the conclusion of program effectiveness (nor of ineffectiveness), in this area.

Finally, the conclusion that the programs *"helped teens in targeted programs to talk more frequently with their parents about issues related to sexuality"* is also questionable. Many of the above methodological concerns apply here as well, with an additional concern that the survey items were analyzed individually (see section III.d page 10), where the appropriate procedure would have been to create scale scores by combining the items. Putting all of these concerns aside, we see that for the programs that submitted cross-sectional data there were no significant differences for MS or HS students, and for elementary schools there was one significant difference, but in the undesired direction. For programs with matched data, all items increased from pre to post test, and most are found significant, but again these results appear erroneous based on the inappropriate use of number of students rather than number of cases in computing p-values. Ignoring the significant tests and just looking at the item means, and imagining scales that could have been created from these items, it does appear, impressionistically, that in this subset of the sample the programs overall might have had a small positive effect, but in fairness any consideration about this would have to be presented in balance with the opposite finding in the cross-sectional group. Further, there are several plausible rival hypotheses other than program effects to explain an increase over time in talking to parents about sex among students with matched data – the most obvious being that as students get older they are more likely to raise questions about sex because they are more likely to engage in sex. This hypothesis would be consistent with finding effects in the matched data sub-sample while not finding these same effects in the cross-sectional data sub-sample; whereas the hypothesis of program effects would not.

One final overall concern I will mention is in regard to the use of effect sizes in various tables. These are impossible to evaluate because the standard deviations that were used in their calculation are not provided. Confidence intervals are traditionally provided around effect sizes but are not provided here. I also noticed that the definition and example of

effect sizes provided in the glossary are incorrect. Yet even if effect sizes were calculated correctly with sufficient supportive information provided to demonstrate that they were, they still would be of questionable utility because they are based on inappropriately pooled data.

That's my quick overview reaction, which turned out not so quick as intended. I hope this is helpful and constructive, although I'm not sure how useful this will be now that the full evaluation report has already been released to the Legislature, the public distribution version is in press, and in any case further program funding is highly unlikely in the current budget crisis. If nothing else, it might be useful to the Department to know in advance what criticisms the report might be open to prior to its public release, and to frame any future communications that the Department makes regarding this evaluation.

Best regards, Norm

Norm Constantine, Ph.D.
Director, Center for Research on Adolescent Health and Development
Public Health Institute
2001 Addison Street, 2nd Floor
Berkeley, CA 94704
Phone: 925.284.8118 Fax: 925.284.8107
nconstantine@phi.org

cc: Wade Brynelson, Assistant Superintendent

PS: After drafting the above letter, I happened to come across the following statement in the Mathematica (2002) federally-funded abstinence-only national evaluation interim report that I am reviewing for our Wellness Foundation project:

> *One implication of the variation in program interventions and services is that it is not possible to reach a single judgment about the efficacy of abstinence education. Such a judgment would only be possible if there were a single, well-defined intervention, one that could vary in its "dosage" across sites but is similar in nature across all sites. In the case of the Section 510 abstinence education programs, however, the interventions and services vary considerably across program sites and sometimes even within a program site. In the absence of definitive evidence on the efficacy of a specific abstinence education approach, this variation is a natural result of the funding opportunities available through Title V Section 510. In addition, the variation in the abstinence education programs provides the opportunity to learn about the effectiveness of different programmatic strategies. (p. 27)*

I think you will see how this reinforces some of my strongest concerns about the TPPGP evaluation conclusions. The Mathematic study was generously funded, had the luxury of several years of planning, randomization to intervention and control conditions within each site, strict quality control on data collection and cleaning, etc. -- all the things that TPPGP didn't have. And further, they decided to focus their resources by selecting only 11 of the several hundred operating programs to evaluate, and only 5 of these 11 for which to conduct impact evaluation, as opposed to process evaluations. Yet they conclude that it is impossible to reach a single judgment about the efficacy of the federal initiative they are evaluating. In other words, about this question, *the results are inconclusive.*