COALITION FOR AN ETHICAL PSYCHOLOGY

human rights * ethics * social justice

www.ethicalpsychology.org

DOES COMPREHENSIVE SOLDIER FITNESS WORK? CSF RESEARCH FAILS THE TEST

ROY EIDELSON
Eidelson Consulting
reidelson@eidelsonconsulting.com

STEPHEN SOLDZ
Boston Graduate School of Psychoanalysis
ssoldz@bgsp.edu

COALITION FOR AN ETHICAL PSYCHOLOGY
WORKING PAPER NUMBER 1
MAY 2012

ABOUT THE AUTHORS

Roy Eidelson, PhD, is a clinical psychologist and the president of Eidelson Consulting, where he conducts research, writes, and consults on the role of psychological issues in political, organizational, and group conflict settings. He is a past president of Psychologists for Social Responsibility, associate director of the Solomon Asch Center for Study of Ethnopolitical Conflict at Bryn Mawr College, and a member of the Coalition for an Ethical Psychology. The former executive director of the Solomon Asch Center at the University of Pennsylvania, Eidelson is the author of numerous scholarly articles on a range of topics in peer-reviewed journals, including intergroup conflict; group identification; anxiety and depression; interpersonal and marital satisfaction; and agent-based computer simulations of political instability.

Stephen Soldz, PhD, is a psychologist and researcher in Boston. He is Director of the Center for Research, Evaluation, and Program Development at the Boston Graduate School of Psychoanalysis. Soldz is especially interested in the intersection of research methodology with clinical and policy efforts. He is an Associate Editor of the *Journal of Research Practice* and on the editorial board of *Psychotherapy Research* and *Politics and Psychotherapy International*. He coedited *Reconciling Empirical Knowledge and Clinical Experience: The Art and Science of Psychotherapy* (APA Books, 2000). He is a member of the Coalition for an Ethical Psychology and is Past-President of Psychologists for Social Responsibility (PsySR). Soldz is a consultant to Physicians for Human Rights and co-authored their report *Experiments in Torture: Human Subject Research and Experimentation in the "Enhanced" Interrogation Program*. He also served as a consultant on several Guantánamo trials.

ABOUT THE COALITION FOR AN ETHICAL PSYCHOLOGY

The **Coalition for an Ethical Psychology** is a group of psychologists dedicated to putting psychology on a firm ethical foundation in support of social justice and human rights. The Coalition works to assure the independence of psychological ethics from government and other vested interests, and has taken a lead role in efforts to remove psychologists from involvement in the abusive and torturous interrogation and detention of national security prisoners. The Coalition is currently spearheading a petition campaign calling for the official annulment of the American Psychological Association's Report of the Presidential Task Force on Psychological Ethics and National Security (the PENS Report), the defining document endorsing psychologists' engagement in detainee interrogations.

DOES COMPREHENSIVE SOLDIER FITNESS WORK? CSF RESEARCH FAILS THE TEST

EXECUTIVE SUMMARY

Comprehensive Soldier Fitness (CSF) is a mandatory resiliency program for all U.S. soldiers that was first introduced in 2009. CSF proponents have identified the program as an urgently-needed response to increasing rates of post-traumatic stress disorder (PTSD), suicide, and other adverse psychological reactions among soldiers exposed to combat in Iraq, Afghanistan, and elsewhere.

Since its inception, CSF has been the target of numerous criticisms. Key concerns include questions about the empirical foundations underlying the program's rapid development and implementation; indications that CSF is a research study involuntarily imposed without appropriate protections for participants; the worry that CSF distracts attention from the need to directly address the adverse effects of multiple and lengthy deployments and high levels of combat exposure; potential negative effects of CSF that have not been carefully considered or monitored; concerns that the "spirituality" component inappropriately promotes religion; ethical questions posed by efforts to build "indomitable" soldiers; issues surrounding the $31 million no-bid contract awarded to psychologist Martin Seligman's positive psychology center at the University of Pennsylvania for CSF development; and the uncritical embrace and promotion of CSF by the American Psychological Association, the world's largest professional association of psychologists.

In this report we highlight and examine a new cause for concern: *CSF evaluation research appears to be deeply flawed and recent claims that the program "works" appear to be gross misrepresentations of the data.* We focus on the report released by CSF researchers last December, titled "*Report #3: Longitudinal Analysis of the Impact of Master Resilience Training on Self-Reported Resilience and Psychological Health Data*." This report is the first to longitudinally assess the central Master Resilience Training component of CSF. After an intensive 10-day training course, each new Trainer is placed in an Army unit. Trainers are charged with equipping fellow soldiers with thinking skills and strategies intended to help them handle the physical and psychological challenges of military life.

With the release of Report #3, CSF researchers asserted: "There is now sound scientific evidence that Comprehensive Soldier Fitness improves the resilience and psychological health of Soldiers." The Army News Service quickly and broadly disseminated the "news": "The Master Resilience Training aspect of Comprehensive Soldier Fitness is working well. That's the conclusion of an Army report, released last month, covering a 15-month period of statistical evaluation."

This overly enthusiastic CSF promotional campaign continues the worrisome and counterproductive history of hyping that began with the program's initial development and roll-out. Inflated expectations have plagued CSF from the start, beginning with exaggerated claims associated with Seligman's Penn Resiliency Program (PRP), the foundation for CSF's Master Resilience Training. A comprehensive meta-analysis of PRP studies concluded that conflicting and inconsistent research findings make it "difficult to give an overall appraisal of the program's effectiveness." Nevertheless, PRP became the basis of the CSF program.

Key Shortcomings of CSF Research Report #3

The claim that CSF's Report #3 provides good evidence that Master Resilience Training is "working well" does not withstand careful scrutiny. The report suffers from multiple inadequacies, including problems with methodology, data analysis, and the interpretation of findings. In our detailed review we examine (1) the failure to measure the important outcomes of PTSD, depression, or other psychological disorders despite the availability of validated measures for doing so, (2) the flawed research design that fails to control for important confounding variables, (3) significant problems with the method of data analysis, (4) the failure to acknowledge plausible risks of the CSF intervention, and (5) other issues of concern.

Report #3 claims that CSF "works" based solely on the Global Assessment Tool (GAT), a 105-item self-report inventory developed for CSF. The GAT does *not* include any validated measures that assess PTSD, depression, suicidality, or other major psychological disorders, even though preventing these disorders is a key goal of the CSF program

and even though such measures are readily available. Thus far there is little evidence that improvement over time in soldiers' GAT scores produces any reduction in the incidence or likelihood of significant psychological distress or other important behavioral health outcomes.

Rather than using the stronger randomized controlled trial research design, Report #3 researchers instead adopted a weaker quasi-experimental approach by choosing which units would include a Master Resilience Trainer; little information is provided as to how these choices were made. Whenever such non-random assignment procedures are used to select groups for comparison, major threats to validity become a serious concern. Among these threats are pre-existing differences between the two groups, as well as the presence of significant confounding variables that might explain between-group differences. A careful reading of Report #3 reveals that the treatment and "control" groups are *not* comparable and that multiple confounds exist. Most notably, half of the soldiers who received CSF training were *deployed* during that time, whereas soldiers who did not receive the training tended to be *non-deployed*. As a result, deployment status could plausibly be more important than CSF training in determining changes in soldiers' GAT scores.

Because soldiers are "clustered" in units, the research design of Report #3 involves data that are not statistically independent. Appropriate analysis of CSF GAT data therefore requires recognition of this statistical nonindependence; even small violations of independence can have very large effects on the accuracy of statistical analyses. But the presence of clustered data is ignored in the most important analyses in Report #3. Until their data have been re-analyzed using the correct techniques, there is little reason to have any confidence in the researchers' reported findings of positive program effects.

While overstating evidence of CSF's effectiveness, Report #3 avoids any serious analysis or discussion of the potential risks of the program. There is no acknowledgment that universal resilience-building interventions, like other types of prevention programs, have a mixed track record, and that unanticipated adverse effects are not uncommon. Those who have investigated potential harm in resilience-building interventions like CSF highlight several dangers. Program participants may subsequently take greater risks if they think they have received some form of preventive protection. Participants may suffer from even greater stigma and shame, perhaps interfering with help-seeking, if after training they fail to effectively handle an adverse event. And the strategies taught may disrupt the participants' prior effective coping strategies. Most people "naturally" respond in a resilient manner when exposed to potentially traumatic events. It cannot be assumed that resilience training will be more helpful than harmful to these individuals.

Conclusions and Recommendations

Certainly the psychological health of our nation's soldiers, and of all citizens, should be a top priority. As a country we must commit ourselves to addressing the alarming rates of PTSD, suicide, and other serious behavioral and emotional difficulties among our troops, especially those repeatedly exposed to the horrors of combat and war. But it is simply wrong at this time to present CSF as part of a solution, because to date there is no solid empirical evidence demonstrating that the program accomplishes any of these lofty goals. Instead, the CSF researchers have examined variables of far less consequence and their methodological approach is riddled with problems – and yet they have broadcast their findings as newsworthy and seemingly deserving of celebration.

It is not hard to imagine the tremendous pressures faced by those responsible for addressing and protecting the psychological health of the men and women who serve in our military. We recognize and admire the dedicated work of so many toward this goal. But in the search for answers, nobody benefits from research that, inadvertently or not, misrepresents the current state of knowledge and accomplishment in this arena. Based on our careful review of Report #3, we believe that the leadership of the Army's Comprehensive Soldier Fitness program must take corrective action. They should give serious consideration to officially retracting the report in its entirety. At a minimum, they should issue an unambiguous and widely disseminated statement acknowledging that the report is seriously flawed and that, as a result, the verdict is still out as to whether CSF actually "works."

DOES COMPREHENSIVE SOLDIER FITNESS WORK? CSF RESEARCH FAILS THE TEST

ROY EIDELSON AND STEPHEN SOLDZ

Using "positive psychology" as its foundation, Comprehensive Soldier Fitness (CSF) is a new and controversial mandatory resiliency program for all U.S. soldiers, first introduced in 2009. In written materials and in communications with Congress, CSF proponents have repeatedly identified the program as an urgently-needed response to increasing rates of post-traumatic stress disorder (PTSD), suicide, and other adverse psychological reactions among soldiers exposed to combat in Iraq, Afghanistan, and elsewhere. Indeed, this urgency has been emphasized as the justification for forgoing the pilot testing that would otherwise be standard for an intervention of this magnitude.

Since its inception, CSF has been the target of numerous criticisms from psychologists and others, including an article, "The Dark Side of Comprehensive Soldier Fitness," we co-authored with colleague Marc Pilisuk last year, a series of critical comments published in the October 2011 issue of the *American Psychologist*, and criticism voiced on *PBS NewsHour*. Concerns raised by critics span a wide range of significant issues: the questionable empirical evidence behind the rapid creation and implementation of CSF; indications that CSF is actually a research study involuntarily imposed upon troops without appropriate protections such as independent ethical review by an institutional review board (IRB) and informed consent; the possibility that CSF may distract attention from addressing the documented adverse effects of multiple and lengthy deployments and high levels of combat exposure; potential negative effects of CSF, common in prevention programs, that have not been carefully considered or monitored, posing the risk of harm to participating soldiers, their families, or civilians in areas where they are deployed; concerns as to whether the "spirituality" component of CSF is inappropriately promoting religion; the insufficient examination of ethical questions posed by efforts to build "indomitable" soldiers; issues concerning the awarding of a $31 million no-bid contract to psychologist Martin Seligman's positive psychology center at the University of Pennsylvania for CSF development; and the seemingly unquestioning embrace of CSF by the American Psychological Association (of which Seligman is a past president), the world's largest professional association of psychologists.

Last year, the concerns that we and other critics had raised led to Congressional inquiries regarding the CSF program. Also in response to these concerns, this past February CSF Directors and research staff invited several of us to meet with them in Washington, D.C, for a forthright discussion about CSF. That meeting focused primarily on questions regarding the effectiveness of CSF in preventing adverse psychological consequences from combat, ethical issues surrounding the program's development and implementation, and concerns about the design and conduct of CSF evaluation research.

In this report we focus on a very serious new concern about CSF, unanticipated at the time of our earlier critique: *the questionable quality of the research being conducted to evaluate the program's effectiveness.* In particular, last December CSF researchers released a report titled "Report #3: Longitudinal Analysis of the Impact of Master Resilience Training on Self-Reported Resilience and Psychological Health Data" (hereafter we will refer to it simply as "Report #3"). In our detailed review of this report, we will discuss (1) the researchers' failure to measure the important outcomes of PTSD, depression, or other psychological disorders despite the availability of validated measures for doing so, (2) the flawed research design that fails to control for important confounding variables, (3) significant problems with the method of data analysis, (4) the researchers' failure to acknowledge plausible risks of harm from the CSF intervention, and (5) other related issues of concern.

Report #3 is the first to longitudinally assess the key Master Resilience Training component of CSF (two previous CSF research reports, released in February and April of 2011, were limited to examining the validity of the program's core assessment instrument, the Global Assessment Tool, or GAT). After undergoing an intensive 10-day training course, each newly-minted "Master Resilience Trainer" is placed in an Army unit. Trainers are charged with equipping fellow soldiers with thinking skills and strategies intended to help them more effectively handle the

physical and psychological challenges of military life, including, most especially, combat operations.

Report #3 presents results comparing soldiers who have received trainings led by Master Resilience Trainers with soldiers who have not. On the report's first page, the CSF researchers boldly assert: "There is now sound scientific evidence that Comprehensive Soldier Fitness improves the resilience and psychological health of Soldiers." And, in a prefatory statement to the report, Army Vice Chief of Staff General Peter Chiarelli writes: "I and other Army senior leaders are often asked if it [CSF] really works – if it actually makes Soldiers more resilient and psychologically healthier. I believe the answer is yes."

Not surprisingly, the Army News Service quickly and broadly disseminated the "good news" this way: "The Master Resilience Training aspect of Comprehensive Soldier Fitness is working well. That's the conclusion of an Army report, released last month, covering a 15-month period of statistical evaluation." This summary has appeared on dozens of websites, including the official websites of the U.S. Army, the CSF program, the Army National Guard, the U.S. Army magazine *Soldiers*, and the *Fort Hood Sentinel.*

Unfortunately, as we will describe here, the claim that CSF's Report #3 provides good evidence that Master Resilience Training is "working well" simply does not stand up to careful scrutiny. This report suffers from multiple flaws, including important issues related to methodology, data analysis, and the interpretation of findings. Equally troubling, the enthusiastic promotional campaign by CSF researchers and other proponents continues the worrisome and counterproductive history of overhyping CSF that began with the $140 million program's initial development and roll-out just a few years ago. *We strongly believe that the CSF leadership should retract Report #3 or, at a minimum, issue a clear, public correction to the record.* The challenges facing our troops are steep enough, without adding to their burdens by promoting unsubstantiated claims while discounting the potential risks from this experimental intervention.

Troubling Background to Research Report #3

Any legitimate and objective evaluation of CSF must recognize that its proponents and spokespersons have repeatedly emphasized that a critical goal of the program is to reduce major adverse psychological reactions among soldiers, including PTSD and suicide. For example, according to Seligman's own account in *Flourish*, in calling for mandatory CSF training for the entire Army (without pilot testing) General George Casey told him and CSF Director Brigadier General Rhonda Cornum, "We are ready to bet it will prevent depression, anxiety, and PTSD." Similarly, in his testimony to the Senate Committee on Appropriations, Casey explained that the CSF program was instituted "to give the soldiers and family members and civilians the skills they need on the front end to be more resilient and to stay away from suicide to begin with." And in a January 2011 special issue of the *American Psychologist* devoted entirely to promoting CSF, Seligman, Cornum, and Michael Matthews described the goal of CSF as "to increase the number of soldiers who complete combat tours without pathology, and to decrease the number of soldiers who develop stress pathologies."

These are worthy aims, but they reflect very high expectations similar to the inflated expectations that have plagued the CSF program from its inception, beginning with the over-hyping of Seligman's "positive psychology"-based Penn Resiliency Program (PRP) as the foundation for CSF's Master Resilience Training. PRP is a group-based intervention program primarily designed to prevent depressive symptoms in school-age children. The leap to modified applications appropriate for soldiers who may be facing life-threatening combat is obviously a large and uncertain one.

Skepticism would therefore be reasonable even if PRP were convincingly effective in its standard school settings. But despite the enthusiastic claims from some, the research evidence on PRP is mixed. Indeed, this is the clear conclusion of a comprehensive meta-analysis of PRP studies conducted by psychologists Steven Brunwasser, Jane Gillham, and Eric Kim in 2009. These investigators concluded that conflicting and inconsistent research findings on PRP make it "difficult to give an overall appraisal of the program's effectiveness." They noted that PRP did not significantly reduce the risk for depressive disorders among any subgroups examined, and that there was no evidence that PRP is superior to active control conditions, such as alternative prevention programs. These authors also emphasized that further

research is needed to determine whether PRP yields practical, meaningful benefits, especially when delivered under real-world conditions.

These authors' scientific assessment of PRP is certainly a warning flag, and their key reservations undercut any judgment that PRP was ready-made for our soldiers in harm's way. This is one reason it is surprising that the Army decided to award a *no-bid* contract in excess of $30 million to the University of Pennsylvania and Seligman's Penn Resiliency Program to develop CSF's Master Resilience Training program. The document authorizing this non-competitive contract (titled "Justification Review Document for Other Than Full and Open Competition – Control No: 09-532") presented PRP in language strikingly different from the cautious evaluation of Brunwasser et al.: "PRP is the only established, broadly effective, evidence-based, train the trainer program currently available which meets the Army's minimum needs" and "The long term outcomes of the PRP have been examined in over 15 well documented studies. The results of the studies have concluded that significant positive effects are sustained and performance of participants is generally improved."

The uncritical praise for PRP has appeared elsewhere as well, including in the January 2011 special issue of the *American Psychologist* edited by Seligman and Matthews. For example, Seligman, his PRP colleague Karen Reivich, and Colonel Sharon McBride of CSF wrote that "Taken together, these findings demonstrate that the skills taught in the PRP lead to significant, measurable positive changes in youth. The preventive effects of the PRP on depression and anxiety are relevant to one of the aims of the MRT course, preventing post-traumatic stress disorder (PTSD)…" At this point, we are left to wonder to what extent such unwarranted enthusiasm for PRP influenced General Casey's decision to forgo a standard period of controlled pilot testing on a limited number of volunteer soldiers – and to instead opt for a massive experimental intervention, without informed consent and other standard research protections, imposed on the entire Army.

This background is worth keeping in mind as we turn now to our specific concerns about CSF Report #3. The report evaluated the PRP-based Master Resilience Trainer component of CSF. Eight Brigade Combat Teams participated in the research, over a 15-month period. Trainers who had completed a 10-day CSF course at the University of Pennsylvania were assigned to four of these teams; this group represented the "Treatment" condition. The other four teams did not have any Trainers assigned to them and comprised the "Control" condition. All soldiers completed the CSF self-report measure of resilience and psychological health – the Global Assessment Tool (GAT) – three times over the course of the study (at "Baseline," then nine months later at "Time 1," and finally six months later at "Time 2"). The Treatment and Control groups were compared in regard to changes in the soldiers' GAT scores from Time 1 to Time 2 as the basis for assessing the effectiveness of the Trainers' interventions.

Failure to Assess the Key Variables of PTSD and Depression

As noted above, CSF training has been heavily promoted as an urgently-needed intervention to reduce the likelihood that U.S. soldiers will commit suicide or experience serious adverse psychological reactions to combat, such as PTSD. However, despite these priorities, Report #3 claims that CSF "works" based solely on the Global Assessment Tool (GAT), a 105-item self-report inventory or questionnaire, with each item answered on a 5-point scale. Although the complete GAT is not publicly available, sample items include the following: "When something stresses me out, I try and solve the problem," "I usually keep my emotions to myself," "When bad things happen to me, I expect more bad things to happen," "I would choose my current work again if I had the chance," and "My life has lasting meaning."

Most of the GAT items measure the "Emotional Fitness" dimension, and almost a third of these items comprise one specific subscale, called "Character." Most importantly, the GAT does *not* include any validated measures that assess PTSD, depression, suicidality, or other major psychological disorders, even though preventing these disorders is a key goal of the CSF program and even though such measures are readily available. It is therefore very troubling that Report #3 touts the "solid evidence" purportedly showing that Master Resilience Training skills are having "a positive effect on Soldier-reported resilience and psychological health." Even worse, the researchers make

the claim, without any substantiation, that "tremendous benefits for the entire Army" can accrue from even small increases in soldiers' resiliency and psychological health. Conceivably that might be true – but there is no evidence provided to support this key assertion.

Indeed, thus far there is little evidence that improvement over time in soldiers' GAT scores (e.g., as a result of CSF training) produces any reduction in the incidence or likelihood of significant psychological distress or other important behavioral health outcomes. To emphasize this point with a simple example, imagine that soldiers who score lower on the GAT tend to be more depressed. This relationship would *not* serve as evidence that increasing a soldier's GAT score is an effective way to reduce his or her depression. It might have no effect on depression at all. For example, soldiers may learn the desired "healthy" responses from going through the training or from repeatedly answering the questionnaire. As is often said, correlation does not prove causality.

The two earlier CSF research reports did provide evidence that higher GAT scores were correlated with higher functioning in several domains. But these studies were not longitudinal. They did not demonstrate that *changes* in GAT scores are associated with *changes* in functional mental health outcomes (for example, differences in GAT scores among soldiers could primarily reflect relatively stable differences in temperament that themselves are associated with life functioning). Producing these latter changes in behavioral outcomes is the central – and so far unassessed – goal of the entire CSF program.

Flawed Research Design Fails to Control for Confounding Variables

The best scientific way to convincingly demonstrate that CSF "works" would be to compare changes in a group of soldiers who receive the training with a comparable group of soldiers who do not. Participants are assessed beforehand and then again after the intervention is completed. But for the post-intervention comparison of the treatment and control groups to be meaningful, it is essential that the two conditions be as similar as possible in all important ways other than the intervention itself. Otherwise any differences that are found could plausibly be explained by these confounding variables. For example, it would obviously be nonsensical to assess the effectiveness of a reading comprehension course by comparing a group that received the training with one that did not, but where only the members of the training group were allowed to use eye glasses when needed.

The gold standard method to assure comparability of the intervention and control groups would be through random assignment of those soldiers who receive the CSF training and those who do not. When the sample is large enough (and certain other conditions are met), random assignment assures comparability. Designs without random assignment – that is, where non-random comparison groups are created by some other procedure – are generally considered weaker than those with random assignment. These designs are called "quasi-experimental" in the research literature in order to call attention to their potential weaknesses. Studies based on quasi-experimental designs traditionally explicitly discuss how these weaknesses – called "threats to validity" – may have affected their results.

For Report #3, the researchers decided on practical grounds that random assignment of soldiers was not feasible. They therefore adopted a weaker quasi-experimental approach by selecting four Brigade Combat Teams to receive CSF training from Master Resilience Trainers and another four Teams to constitute the comparison group that did not receive training during the study. Regrettably, Report #3 provides little information on how these Teams were selected, information that is central to assessing the trustworthiness of research comparing their outcomes. The report also contains almost no discussion of potential threats to validity and their potential impact on its findings.

Whenever non-random assignment procedures are used to select groups for comparison, major threats to validity become a serious concern. Among these threats are any pre-existing differences between the two groups, as well as the presence of significant confounding variables during the intervention that might explain between-group differences in outcome measure changes. Unfortunately, a careful reading of Research Report #3 reveals that the treatment and "control" groups are not comparable and that multiple confounds exist. One stands out as especially problematic: approximately half of the soldiers who received the CSF training were

deployed during that time period, whereas the soldiers who did not receive the training (i.e., the comparison group) tended to be *non-deployed* (see Figure 1 in the report). In short, deployment status could plausibly be far more important than resilience training in determining changes in a soldier's GAT scores. After all, it is easy to imagine how different life must be for soldiers assigned to a military base in the U.S. compared to their counterparts deployed to the rugged mountains of Afghanistan. It simply does not make sense to assume that differences in soldiers' self-reported feelings, thoughts, and mood are the result of whether or not they have received CSF training, rather than whether or not they are deployed. Yet this assumption is a foundation of Report #3.

It should be noted that it is not clear *a priori* whether deployment would tend to increase or decrease self-reported resilience and psychological health as measured by the GAT. For example, being deployed may influence friendships, unit cohesiveness, and one's personal sense of efficacy. Similarly, the anticipation of deployment may heighten feelings of excitement *or* dread; and the anticipated return home from deployment may produce thoughts related to either relief *or* regret, or both. Again, with such deployment differences between the treatment group and the comparison group, it is likely impossible to draw meaningful conclusions about the CSF training itself.

Flawed Data Analysis Procedures

Research data need to be carefully analyzed to determine whether there is evidence of differences between groups or changes over time. Trustworthy data analysis requires that the data meet the underlying assumptions of the statistical models that are employed. For the traditional data analytic techniques used in most of the analyses presented in Report #3, a basic assumption is that each soldier's data are "independent" from data provided by other soldiers. In other words, one soldier's GAT scores must not be influenced by or related to the GAT scores of another soldier.

Data failing to meet this independence assumption are often called "clustered data" because clusters of observations are potentially more similar in characteristics than are randomly chosen observations. A classic case of non-independence is found in educational research where students in the same classroom are more likely to have similar test scores than a group of randomly chosen students. This is true for several possible reasons: the students may have been assigned to the class based upon comparable abilities in the subject matter; they may have chosen to be together in this particular class; they interact with each other in the learning process; they may study together; and they have the same teacher, with his or her specific style of teaching. Educational data are therefore said to be clustered in classrooms. In the same way, students from one school are likely to be more similar in academic achievement than are students from different schools, due to clustering on the basis of socioeconomic and other factors.

Even small violations of this independence assumption can have very large effects on the accuracy of statistical analyses. As a result, in situations where nonindependence is likely, researchers should adopt special statistical techniques designed for analyzing clustered data. To ignore such clustering usually produces biased results and often a greater likelihood of reporting "statistically significant" findings (e.g., differences between groups) that disappear when the data are analyzed correctly.

The design of the CSF Report #3 study involves nonindependent data similar to the school setting example described above. The soldiers clustered in units with the same Master Resilience Trainer are analogous to students in classrooms. Soldiers in the same unit are likely to be more similar in GAT scores for several reasons: they may have had similar experiences; they may influence each other's attitudes and coping patterns; and like students in classrooms, soldiers in the same unit interact with each other on a daily basis and have the same Trainer. Appropriate analysis of the CSF data therefore requires careful consideration of the potential effects of clustering. However, even though the authors of Report #3 are aware of the issues surrounding clustered data and the need to use specialized statistical techniques (indeed, they even examine clustering in one of their subsidiary analyses), they inappropriately ignore this clustering in their most important analyses of the GAT data.

The authors of the report indirectly defend their failure to utilize the appropriate techniques for clustered data by incorrectly claiming that clustering will have only negligible effects on their

findings. However, as we demonstrate in our Technical Appendix, the researchers' own data, as well as statistical authorities they cite, suggest that the clustering may have large effects on the reliability of their results. In particular, by ignoring the clustering the authors may have incorrectly identified and interpreted key findings as "statistically significant." Until their data have been re-analyzed using the correct techniques, there is little reason to have any confidence in the researchers' reported findings of positive program effects.

Failure to Acknowledge Plausible Risks of Adverse Effects

Just as Report #3 overstates evidence of CSF's effectiveness, it also avoids any serious analysis or discussion of the potential risks of the program. In particular, there is no acknowledgment that universal resilience-building interventions have a mixed track record, and that unanticipated adverse effects are not uncommon. As we noted in our earlier critique, some criminal justice prevention programs have been shown to increase future offending, and some substance abuse prevention programs have failed to reduce – and in some cases have even *increased* – abuse-related behaviors.

Those who have carefully investigated potential harm in resilience-building interventions like CSF highlight several dangers (see the 2011 review by George Bonanno, Maren Westphal, and Anthony Mancini for more on this important subject). Program participants may take greater risks than before if they think they have received some form of preventive protection (in the case of CSF, this increased risk-taking could pose dangers to peers or civilians in areas where the unit is deployed). Participants may suffer from even greater stigma and shame, perhaps interfering with help-seeking, if they are unable to effectively handle an adverse event after having received the training. And the strategies taught and recommended may disrupt the participants' already effective prior coping strategies. In regard to this last point, it is important to recognize that most people – upward of two-thirds – "naturally" respond in a resilient manner when exposed to potentially traumatic events (and this figure may be even higher among soldiers). It cannot be assumed without evidence that resilience training will be more helpful than harmful to these individuals.

In thinking about potential harmful effects of an intervention, we should distinguish between two types of harm. In one case, as in other prevention programs that were found to cause harm, they may lead to *worse average functioning* than no intervention. These harms can be detected by traditional statistical analyses. However, some interventions are helpful, or at least not harmful, on average while causing harm to certain individuals or subgroups who receive the intervention. These harms may remain undetected by traditional analyses aimed at detecting mean change. Thus, with medications and medical devices, an adverse impact monitoring system is often used to try and detect sometimes rare, harmful effects in otherwise helpful interventions.

In this context, the statement in Report #3 that there is no evidence that soldiers "get worse" due to the CSF training is highly misleading. It is not even clear how the researchers could have reached this conclusion. Consider the following. Given the minimal overall change from Time 1 to Time 2 on GAT scores at the group level, and recognizing that change is usually randomly distributed with a bell-shaped distribution, there is little doubt that a substantial number of individual soldiers scored *worse* on the GAT's measures of resilience and psychological health *after* the CSF training than before. Such declines cannot be attributed directly to the CSF intervention itself. But in just the same way, CSF training cannot be ruled out as a possible contributor to this worsening for some.

On this point, public health researchers such as Geoffrey Rose have emphasized that there are important risk differences between two types of mass prevention programs. On the one hand, there are those interventions that aim to return individuals to conditions that are more "normal" (e.g., by reducing excessive consumption); the population risks associated with these programs are likely to be quite small, while simultaneously providing substantial benefits to some. On the other hand, mass interventions that focus on adding some form of protective enhancement (e.g., by taking herbal supplements to prevent illness) entail a substantially greater risk to the population, and the benefits are often less certain as well. Standards of evidence documenting safety therefore need to be more stringent in these latter cases, of which CSF is a prime example.

Additional Related Concerns with Report #3

The various problems already highlighted here are more than sufficient reason to reject the view that Report #3 demonstrates that CSF "works." But several additional points merit brief mention as well. First, the researchers' enthusiastic appraisal of CSF is based entirely on a handful of reported GAT score differences between the treatment group and the comparison group of little more than 1% to 2% over a nine-month period. These results are even weaker than they already appear when one considers the minimal associated effect sizes in Report #3 and the experiment-wise error involved in conducting many statistical significance tests without adjusting for the number of analyses conducted. The authors offer reassurances for these weak findings by suggesting that small improvements can have important consequences in the world of prevention. This can certainly be true in the public health domain where, for example, a 3% increase in the number of people who quit smoking can have substantial positive effects on those individuals and on society more broadly. But as noted earlier, Report #3 does *not* assess significant outcomes that correspond to important changes in behavioral health (such as PTSD or suicide rates). Rather it relies entirely on changes in self-reported GAT scores over time – without any evidence that modest changes in GAT scores are associated with meaningful real world effects.

Second, in a footnote the authors of Report #3 acknowledge that one of the central components of the CSF program – its online Comprehensive Resilience Modules – had no impact in promoting soldier resilience and psychological health. The modules are reportedly now undergoing significant revision, but they continue to be highly touted on the CSF website (and elsewhere) as one of the "pillars" of CSF, despite their apparent ineffectiveness. This ineffectiveness is consistent with our view that pilot-testing research should have been a prerequisite for the rollout of CSF – it simply is not a training program of well-documented and established value. Further, the failure of CSF researchers to fully report these negative results raises concerns as to whether they are "cherry-picking" for dissemination those findings they consider supportive of the program.

Third, Report #3 lacks sufficient clarity in regard to the nine-month period between two key time-points, "Baseline" and "Time 1." In fact, no Baseline data are presented or described. As a result, the reader does not know what happened during this period, which is particularly problematic in regard to the Treatment condition. The reader also cannot tell whether GAT score changes were measured between these two time-points for the treatment and comparison groups. One possible explanation for the group differences at Time 1 is that there was already an effect of resilience training and placement during the preceding nine months, even if the program implementation may have been more haphazard than desired during this period. Alternatively, the two groups may have been different from the very beginning (i.e., starting at Baseline). This latter possibility provides additional reason to be very cautious about the analyses claiming to demonstrate a positive effect for the Master Resilience Trainer program.

At our February meeting with CSF directors and researchers, we were told that CSF scores actually *declined* from Baseline to Time 1 in *both* groups – suggesting an overall worsening in psychological functioning over this period. This despite the fact that the Master Resilience Trainers in the treatment group were already trained and were supposed to be developing resilience skills among the soldiers assigned to them. The researchers claimed that the decline was not that surprising due to a lack of clear, published guidance for these Trainers prior to Time 1. Although this may explain the absence of an increase in GAT scores from Baseline to Time 1, it sheds no light on the reasons for a measurable decline. The researchers advised us that these results – and thus these Baseline GAT scores – were omitted from Report #3 due to their "complexity." In any case, this decline is noteworthy and it suggests that there are important factors affecting GAT scores other than the CSF intervention. Further, failing to include the Baseline scores leaves the reader in the dark as to whether the treatment and comparison samples were actually similar when first selected, which is the comparability that really matters. The omission of this important information again raises significant concerns about the process used for selecting which results would be presented in the report.

Fourth, some tables in Report #3 are

confusingly labeled. Thus, Tables B1 and B2 (pp. 49-50) both appear to report the Time 1 means for the Treatment and Control groups, but the values differ from one table to the other. (We thank Sean Phipps for bringing this discrepancy to our attention.) An email from a CSF researcher explained that this discrepancy was because the reported means were not the actual means, but were so-called "estimated marginal means" that resulted from specific statistical analyses, which differed for the two tables. This incomplete or mislabeling in tables, combined with the incomplete reporting of results, makes the report difficult for independent researchers to properly evaluate.

Finally, given how broadly and repeatedly the CSF program has been aggressively promoted within the Army (despite the lack of research evidence), there may well be significant "demand characteristics" operating here. That is, many soldiers likely received advance notice of the program's reported benefits, and some who subsequently completed the GAT may have been influenced by those expectations. Such pressures could represent another potentially important confounding variable.

Conclusions and Recommendations

Based on our careful review of Research Report #3, we believe that the leadership of the Army's Comprehensive Soldier Fitness program must take corrective action. They should give serious consideration to officially retracting the report in its entirety. At a minimum, they should issue an unambiguous and widely disseminated statement acknowledging that the report is seriously flawed and that, as a result, the verdict is still out as to whether CSF actually "works."

In making this recommendation we fully recognize that large-scale evaluation research is an intrinsically difficult undertaking inevitably imperfect in its execution. However, the public that has paid over $100 million for the CSF program and, even more, the one million soldiers who are involuntarily subjected to CSF's resiliency training deserve much better than the misrepresentations of effectiveness aggressively promoted by Report #3.

Certainly, the psychological health of our nation's soldiers, and of all citizens, should be a top priority. As a country we must commit ourselves to addressing the alarming rates of PTSD, suicide, and other serious behavioral and emotional difficulties among our troops, especially those repeatedly exposed to the horrors of combat and war. But it is simply wrong at this time to present CSF as part of a solution, because to date there is no solid empirical evidence demonstrating that the program accomplishes any of these lofty goals. Instead, the CSF researchers have examined variables of far less consequence and their methodological approach is riddled with problems – and yet they have broadcast their findings as newsworthy and seemingly deserving of celebration.

This puzzling reality is quite distressing. We have argued elsewhere that CSF is a massive research project without informed consent, one that should never have been universally implemented prior to careful piloting testing. But even when prevention programs truly do work (and again, such an assessment of CSF is so far unwarranted), it is expected that the researchers will assess and report any weaknesses in their evaluation and any potential harm to participants in the program. Report #3 fails to seriously engage with – or even acknowledge – either of these critical issues.

These scientific shortcomings are all the more troubling given the obvious importance of what is at stake here: soldiers' welfare. It may be comforting to some to assume that, at worst, CSF is merely ineffective. However, we should not settle for such wishful thinking. It is not outlandish to suggest that CSF may negatively impact some soldiers, and unjustified enthusiasm about the program can prove costly in terms of directing attention and funding away from the consideration and development of alternatives that may be far more beneficial for our troops.

It is not hard for us to imagine the tremendous pressures faced by those responsible for addressing and protecting the psychological health of the men and women who serve in our military. We recognize and admire the dedicated work of so many toward this goal. But in the search for answers, nobody benefits from research that, inadvertently or not, misrepresents the current state of knowledge and accomplishment in this arena. For this reason, we believe it is essential that the Comprehensive Soldier Fitness leadership correct the record in regard to their Research Report #3.

Acknowledgments. *We wish to thank James Coyne, John Dyckman, Joachim Krueger, Sean Phipps, James Quick, and Stephen R. Shalom for helpful comments on earlier versions of this report.*

References

Benjamin, M. (2011). "War on Terror" Psychologist Gets Giant No-Bid Contract. http://www.salon.com/2010/10/14/army_contract_seligman/ .

Bonanno, G. A., Westphal, M., & Mancini, A. D. (2011). Resilience to Loss and Potential Trauma. *Annual Review of Clinical Psychology, 7,* 511-535.

Brunwasser, S. M., Gillham, J. E., & Kim, E. S. (2009). A meta-analytic review of the Penn Resiliency Program's effect on depressive symptoms. *Journal of Consulting and Clinical Psychology, 77,* 1042-1054.

Cornum, R., Matthews, M. D., & Seligman, M. E. P. (2011). Comprehensive Soldier Fitness: Building resilience in a challenging institutional context. *American Psychologist, 66*, 4-9.

Dyckman, J. (2011). Exposing the Glosses in Seligman and Fowler's (2011) Straw-Man Arguments. *American Psychologist, 66,* 644-645.

Eidelson, R., Pilisuk, M., & Soldz. S. (2011). The Dark Side of Comprehensive Soldier Fitness. http://www.psychologytoday.com/blog/dangerous-ideas/201103/the-dark-side-comprehensive-soldier-fitness-0.

Eidelson, R., Pilisuk, M., & Soldz. S. (2011). The Dark Side of Comprehensive Soldier Fitness. *American Psychologist, 66,* 643-644.

Krueger, J. (2011). Shock Without Awe. *American Psychologist, 66,* 642-643.

Lester, P. B., Harms, P. D., Bulling, D. J., Herian, M. N., & Spain, S. M. (February 2011). Evaluation of Relationships between Reported Resilience and Soldier Outcomes. Report #1: Negative Outcomes (Suicide, Drug Use and Violent Crime). http://www.dtic.mil/dtic/tr/fulltext/u2/a538618.pdf.

Lester, P. B., Harms, P. D., Bulling, D. J., Herian, M. N., Spain, S. M., & Beal, S. J. (April 2011). Evaluation of Relationships between Reported Resilience and Soldier Outcomes. Report #2: Positive Performance Outcomes in Officers (Promotions, Selections, & Professions). http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA542229.

Lester, P. B., Harms, P. D., Herian, M. N., Krasikova, D. V., & Beal, S. J. (December 2011). The Comprehensive Soldier Fitness Program Evaluation: Report #3: Longitudinal Analysis of the Impact of Master Resilience Training on Self-Reported Resilience and Psychological Health Data. http://dma.wi.gov/dma/news/2012news/csf-tech-report.pdf.

Phipps, S. (2011). Positive Psychology and War: An Oxymoron. *American Psychologist, 66,* 641-642.

Quick, J. C. (2011). Missing: Critical and Skeptical Perspectives on Comprehensive Soldier Fitness. *American Psychologist, 66,* 645.

Reivich, K. J., Seligman, M. E. P., & McBride, S. (2011). Master resilience training in the U.S. Army. *American Psychologist, 66*, 25-34.

Rose, G. (1981). Strategy of prevention: lessons from cardiovascular disease. *British Medical Journal, 282,* 1847-1851.

Sagalyn, D. (2012). Health Experts Question Army Report on Psychological Training. Retrieved from http://www.pbs.org/newshour/updates/military/jan-june12/csf_training_01-02.html.

Seligman, M. E. P. (2011). *Flourish: A Visionary New Understanding of Happiness and Well-Being.* New York: Free Press.

Senate Hearing 111-688: Department of Defense Appropriations for Fiscal Year 2011. http://www.gpo.gov/fdsys/pkg/CHRG-111shrg54962/pdf/CHRG-111shrg54962.pdf.

TECHNICAL APPENDIX

In all analyses in Report #3, the data are treated as if observations are independent, a basic assumption of all traditional statistics (that is, statistics not explicitly taking nonindependence, or "clustering," into account). In only one section, that involving the effects of the Master Resilience Trainers (MRTs), do the researchers even examine the presence of clustering. The section on MRT effects (p. 21) demonstrates that the assumption of independence is violated, in that soldiers in the same MRT unit are more similar to each other on GAT-assessed resilience and psychological health (R/PH) than soldiers selected at random. In particular, Table B7 (p. 55) shows that there was nonindependence in the form of significant Intraclass Correlations (ICCs). The Report #3 authors discuss this issue by stating that these ICCs "ranged from trivial (.001) to very small (.036)" (p. 21), referencing Julian (2001) among other sources. However, they misrepresent what Julian says. He states, "When the magnitude of the intraclass correlations are less than .05 *and the group size is small*, the consequences of ignoring the data dependence within multilevel data structures seems to be negligible" (p. 347, emphasis added).

But group size is *not* small for the Report #3 analyses, and the authors wrongly ignore the important effect of group size in their interpretation of the ICCs. Relatively small ICCs can have *large* effects when the groups are large (here, the number of soldiers assigned to an MRT's unit). The average cluster size in this research report is approximately 100 (4,348/44, see p. 21). Julian's Table 1 (p. 339) demonstrates the very substantial effect of a "small" ICC of .05 when the cluster size is 100. In this case, a nominal alpha of .05 (i.e., an expected 25 rejections out of a sampled trial of 500) in fact becomes a real alpha of .21 (i.e., 104 rejections out of 500). Thus, all significance tests in Report #3, except for the MRT analyses, are likely biased, often to a significant degree, and they cannot be considered accurate. Given that the authors are aware of this issue and correctly used multilevel models in their MRT analyses, it is unclear why they did not use this approach to address the issue of nonindependence throughout the report.

Standard measures of the extent to which statistical results are invalidated by nonindependence are the Design Effect (DEFF) and its square root (DEFT). DEFF is the factor that reduces the sample size to reflect the condition of non-independence. DEFT is the corresponding factor that increases the standard errors and confidence intervals. The problem becomes apparent when examining the formula for DEFF calculation: DEFF = 1 + (n-1) * ICC, where n is the average cluster size. Thus, a "small" ICC combined with a large cluster size can have a large Design Effect. For ICCs of the magnitude reported for the CSF data and group sizes of 100, DEFF (and DEFT) can be sizeable and meaningfully different from a neutral 1.0. Thus, for an ICC of 0.01, DEFF would be 1.99 and DEFT 1.41, while for the largest ICC found in Table B7 of Report #3, 0.36, DEFF is 4.56 and DEFT in 2.14.

Another way of examining the importance of these design effects is via a table in Kreft and de Leeuw (1998, p. 10), originally attributed to Borcikowski, which conveys the observed alpha levels for certain models with a nominal alpha of .05 for various levels of cluster size and ICC. With cluster size of 100, as in Report #3 analyses, alpha would be .17 for an ICC of .01 and .43 for a "very small" alpha of .05. It should also be noted that, as Bliese and Hanges (2004) demonstrate, with some models nonindependence can cause significance tests to be too conservative rather than too liberal. While estimating the effects of clustering on significance tests depends upon the details of the model being examined, these results indicate that it is a serious mistake to ignore the clustering in the CSF data and that the reported significance levels in Report #3 may be seriously inaccurate.

It should be noted that the CSF data may have a second level of clustering because MRT units are clustered in brigades. This possibility should have been explored as well.

References

Bliese, P. D., & Hanges, P. J. (2004). Being both too liberal and too conservative: The perils of treating grouped data as though they were independent. *Organizational Research Methods, 7*, 400-417.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8*, 325-352.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.