



## The Context and Clarification of a Single Study: Response to Commentaries on *One Year Longitudinal Study of the Psychological Effects of Administrative Segregation*

By Maureen O'Keefe, Kelli Klebe, Jeffrey Metzner, Joel Dvoskin, and Jamie Fellner

### Abstract

*In this article, authors and advisory board members of the Colorado study respond to commentaries published at **Corrections & Mental Health** on their 2010 report, **One Year Longitudinal Study of the Psychological Effects of Administrative Segregation**.*

**Keywords:** administrative segregation, Colorado, methodology

We are pleased to contribute *One Year Longitudinal Study of the Psychological Effects of Administrative Segregation* (O'Keefe et al., 2010) to the research literature on solitary confinement and to engage in scientific dialogue regarding research methods, the interpretation of results, and the need for future research. We appreciate the reviewers' thoughtful comments and criticisms reported in this issue of *Corrections & Mental Health* and we thank the National Institute of Corrections for the opportunity to respond to them.

### Background

From the very start, a primary objective of this study was to advance the quality of research methods when conducting research to understand the impact of administrative segregation on prisoners' psychological distress. We recognized that the difficulty of doing research within correctional institutions – let alone research of such a sensitive nature – often prohibits good research design. It is difficult to pass through concrete and mortar walls; administrative protocols are confusing to navigate; data systems are unwieldy and complex; and vigilant institutional review boards are formidable when studying protected populations. At the same time, the researchers also recognized from past experiences that the

environment in Colorado was ripe for this study. Maureen O’Keefe, the study’s Principal Investigator (PI), and Kelli Klebe, the study’s co-Principal Investigator (Co-PI) had collaborated on numerous projects at the Colorado Department of Corrections (CDOC) to know that with the agency’s executive director (Aristedes Zavaras) approval we would receive system-wide support, gain access to inmates, have electronic access to offender data systems, achieve independence and autonomy to conduct our research, and have a commitment to complete the project. There was every reason to believe CDOC’s executive director when he said he wanted to be the first to know if administrative segregation (AS) was causing psychological harm so that the department could adapt its policies, and we expected to have plenty of recommendations to share when the results were done.

It was the focus on improving research methods and ensuring the objectivity and integrity of the research that led the researchers to solicit the involvement of three external forensic, correctional, mental health, and human rights experts; it was also the reason that they agreed to volunteer on the advisory board alongside several Colorado wardens and mental health administrators. Like many of today’s critics who are unfamiliar with the research that has been done in and about the Colorado corrections system, the outside advisors were skeptical of the ability of the researchers to carry out this project. Would Colorado permit this research? Would the researchers be receptive to feedback? Did the researchers have the skills to manage day-to-day research operations? Did they possess the statistical savvy to analyze complex data sets? Would inmates be willing to participate and be forthcoming about their psychological symptoms? Despite their fears – and those of the researchers as well – the outside experts each agreed to take it on. From the first meeting of advisory board members during a blizzard at the famous Stanley Hotel in Estes Park, Colorado, where they ripped the study design apart and put it back together again in the best possible way, to the project’s end with results that shocked *nearly* all, these experts walked every step of the project with the researchers. The reason the advisory board was so successful, even through periodic squabbles of the most opposing viewpoints conceivable and hours-long discussions of research and philosophical perspectives, is because every single member of the advisory board believed in the value of empirical research.

Frankly, the results of the study came as a surprise to the investigators and the external advisory board members. Based on reviews of the literature or having seen inmates in segregation experience intense psychological distress, it was our collective expectation that we would find an identifiable group of inmates with characteristics that made them especially vulnerable to the psychological stressors inherent in solitary confinement.

In retrospect, this project encountered surprisingly few obstacles. Having previously accessed prison inmates and helped outside researchers to do the same, the PI was able to secure support and entry into the institutions for the field researcher. We even conducted a pilot study to test the procedures and responses to our instruments (i.e., whether too difficult, too long, or irrelevant items). Because of the PI’s knowledge of the offender data system, she downloaded data from the department’s data system. However, instrument data collection, data entry, and quality assurance were done by university staff. The PI reviewed data only to the extent necessary to ensure adequate subject accumulation and timely collection of data and to check that the instrument scoring syntax was correctly programmed. We had a field researcher with an amazing penchant to follow the different security and scheduling protocols for each facility, requiring no intervention on the study’s behalf (rare, in our experience). The Co-PI, whose

specialty is statistical analysis and research design, analyzed all data except the baseline comparisons of downloaded institutional data; the PI managed the project to ensure proposed grant timelines were met, including the completion of the final written report. Thus, each member of the team was involved in a way that maximized his or her unique skills, allowing us to achieve our grant goals and to incorporate checks and balances to ensure objectivity.

The first 18 months of the project took more resources than anticipated to build and automate research data systems, to implement training and quality assurance, to ensure steady sample accumulation, to manage data collection during the height of testing (during the middle period of data collection, we were obtaining informed consent from subjects while others were completing their final testing), and to code data and write computer code for data scoring and analyses. After that point, the project became very easy to manage, with little to do while waiting for the field researcher to complete data collection. When National Geographic proposed a documentary on our study, we had nearly complete data through the fourth interval and the Co-PI ran the first analyses. This was our first indication of the direction of the results. By the time the film crew arrived at Colorado State Penitentiary, there were only two offenders there left to complete the final testing session; no study participants were in the documentary and the documentary did not air until all testing was complete.

The researchers secured agency approval and support at the outset, and we had about six CDOC staff on the advisory board (there were more in the beginning but some lost interest or took other positions) who participated in meetings and assisted the research team when we encountered difficulties. Other than that, CDOC management had a hands-off approach, granting the research team complete autonomy. In fact, only the advisory board members and our invited American Psychological Association conference (Klebe, 2010) discussant, Dr. Stuart Grassian, were given the opportunity to review, comment, and influence the final report. (The granting agency also had two anonymous reviewers.) Advisory board members from within CDOC only commented on sections that detailed agency policies or practices; external reviewers provided far more robust, lengthy, and detailed feedback over the course of several revisions.

### **Response to Methodological Issues**

We agree that this study is not flawless, and as such, we acknowledged certain study limitations in the discussion of the full report. However, we would like to address certain criticisms of the current reviewers, with which we may disagree or need to put into context. The reviewers have made several criticisms about our study with which we agree, some of which were also included within our own study limitations section of the full report.

#### *Self-Report Data*

Several reviewers are concerned with the use of paper-and-pencil tests as the major means of collecting data, even suggesting that they were the sole means of data collection. They express several concerns about the data including the unreliability of the self-report data, claims about the field researcher, and response bias due to being studied.

The criticisms of self-report data, of course, apply to virtually all studies of human distress, especially in solitary confinement where behavioral options are limited. The psychological harms alleged by most critics of long-term segregation are related to phenomenological experiences of inmates; and while self-report is not a perfect measure, it is the only way to know about a person's subjective experiences. Thus, it is not surprising that previous researchers, such as Haney (2003) and Grassian (1983) relied exclusively on self-report. However, unlike Haney and Grassian, we did not rely solely on self-report.

*Multiple Data Sources.* Data were also collected from (1) clinician ratings on the *Brief Psychiatric Rating Scale*, which was scored based on observation and interview of an individual by prison mental health staff following standardized protocols, (2) correctional officer ratings of behavior using the *Prison Behavior Rating Scale*, (3) mental health crisis reports, and (4) prison logs of behavioral data and out-of-cell activities. All data are summarized in the full report.

Both clinician and correctional officer ratings failed to show mean negative change over time, which is consistent with the inmate self-report data. Unfortunately the measures by clinicians and correctional officers had lower reliability coefficients than self-report data and also had low correlations to the self-report data. There are multiple reasons why the staff and inmate measures may not correlate, making it difficult to interpret differences in responses. Some possible explanations include that participants may respond differently on pencil-and-paper measures than in clinical interviews, clinicians may minimize psychological symptoms due to not believing the client or scoring biases, under-staffing may not allow clinicians sufficient time to spend with inmates, and correctional officers may not know an inmate well enough to report accurately on some behaviors.

We had hoped to gain more insight into inmates' conditions of confinement as actually experienced, in contrast to formal policy, from prison logs; however, the incompleteness of the logs did not make this possible. There was so much missing data from the prison logs that it was not possible to interpret why data were missing. Possible explanations include failure of staff to complete logs, failure to offer inmates the opportunity to exercise/shower, inmates refusing to leave their cells, etc. Although the researchers had access to these logs, we were careful not to influence their content, we were not responsible for the keeping of the data, and therefore did not have control over their completeness or quality. Thus, we concluded that these data were unusable for the purposes of this study.

There has also been criticism of not using the mental health crisis data; however these data are described in the study report. Crisis events involve any situation that is not a scheduled appointment and requires immediate psychological intervention. Events were coded on a self-harming dimension (ideation, self-harm behavior, suicide attempt) and a psychotic symptom dimension (yes/no for any psychotic symptom); many events involved both a psychotic symptom and a self-harming symptom that should not be construed as two separate events. It is also important to note that we coded psychotic symptoms in a broad, inclusive manner. For instance, if the person making the mental health crisis referral made mention of a hallucination or delusion, the event was coded as having a psychotic symptom even if not observed by a trained clinician and if denied by the offender. We believe that there are difficulties interpreting these data because we do not know what the pre-study incidence rates were for these individuals, whether similar symptoms were addressed during regularly scheduled mental health appointments, and whether such behaviors went undocumented because they occurred without staff's

knowledge. Furthermore, these data do not indicate the reason for the crisis event. For example, a mentally ill inmate in AS threatened suicide because he did not want to be progressed to a specialized treatment program at a lower security prison. In another example, an inmate in AS reported self-harming behavior due to grief over the death of several family members as a result of an automobile accident; it would be faulty to attribute such an event to the conditions of confinement rather than unfortunate life circumstance. These data contain too much ambiguity at this point and we disagree with reviewers and others who claim these are important sources of behavior which we ignore because they appear to be contrary to our conclusions.

Our design included other measures beyond self-report and we expected quality data from these measures, but this did not occur. These problems explain some of the reasons that our conclusions focused on the self report data over official record data and staff data. The reliability coefficients that emerged from the self-report data are another reason we placed a high degree of confidence in these measures.

*Reliability and Validity.* The majority of the self-report instruments used in this study are standardized measures with established reliability and validity for use in the general population. Several measures have been used with prison populations and we cite this literature in Appendix B of our report. We provided information about how our data are similar and dissimilar to other research. The measures were selected to assess the SHU Syndrome symptoms (Grassian, 1983)

We used multiple measures for each psychological or cognitive construct in order to assess consistency across similar measures as well as tap into the multidimensional nature of the constructs of interest. We provided reliability estimates of the tests completed by study participants. The measures demonstrated internal consistency in responding, stability over time, and similar results as normative data. Importantly, individuals tended to be consistent across measures of the same construct and across time (Table 8, p. 39 and Appendix B, pp. 127 - 137 ), indicating reliability in responses. The strong reliability estimates and consistency in responding to items within a measure, across similar measures of the same domain, and across time on the same instruments and constructs indicate low measurement error in responses.

Strong positive relationships between measures of the same construct and weaker relationships between different constructs provided evidence for convergent validity. Additionally, the study groups tended to rank order themselves on mean scores in a way that makes sense (e.g., mentally ill inmates scored higher than those without mental illness), another measure of convergent validity.

We would like to clarify our use and conclusions regarding the SIMS data based on comments by Shalev and Lloyd. They state “We are told that most of the participants (85%) had an elevated score on at least one of the five subscales ... In fact Table 7 indicates that between half and two thirds of the whole sample returned an inconsistent profile on each scale.” Table 7 (p. 38) in the report gives the percentage of people who had scores above the suggested cutoffs for malingering on each of the SIMS subscales by group for each interval; we also computed a score for each individual indicating whether or not he was elevated on at least one subscale. Nearly 85% had at least one elevated score (individual group percentages ranged between 49% and 97%), but most people did not have multiple elevations. Subscales on the SIMS may not be correlated if persons have a true mental health issue as opposed to

overall malingering (where you might expect consistently high scores across SIMS subscales). Because the SIMS manual indicated that high scores could mean either malingering or true psychopathology, we decided not to use the SIMS as a malingering measure. We believe that Shalev and Lloyd have misinterpreted the statistics, although our general findings are in agreement with their conclusion that study participants demonstrated a high degree of distress on psychological variables.

*Participant Disclosure.* There are several comments regarding whether or not participants were willing to tell us the truth and whether they were underestimating their responses in order to show that they had adapted well or to hide vulnerabilities. It is possible that although participants respond in consistent, reliable ways, they were not responding *truthfully* concerning their psychological well-being. Although this is possible, there are several trends in the data that lead us to question this criticism: (1) participants scored high on the measures relative to normative data, which is inconsistent with the criticism that they are underreporting; (2) the study groups tended to rank order themselves on mean scores in a way that makes sense (e.g., inmates placed in the mental health hospital tended to have highest scores); (3) participants responded in consistent ways over multiple measures of the same construct and over multiple assessment periods, which would be difficult to maintain over one year if they were not speaking to some internal truth of their own. This being said, we agree that general under-reporting is possible, but do not see any reason to expect or believe that underreporting would be higher for segregation inmates than for the various comparison groups.

*Researcher Effects.* Another common criticism of the study was that collecting data from inmates increased the number and quality of interactions that participants had, which may have lowered the impact of solitary confinement. Although this is a possible explanation for the findings, it would be true of any research using any type of self-report data from participants (whether interviews or paper-pencil tests). However, we would expect this impact to be less with paper-and-pencil tests than with interviews because the quality of interactions are more limited when giving test instructions compared to diagnostic interviews. Interviews have the additional problems of researcher bias and leading questions. We agree that quality interview data provide the opportunity to understand the context of participants' responses. However, in our study, we were addressing the criticism that most studies have not used standardized data with known psychometric properties.

There have been several complaints about our use of a field researcher without a doctorate degree (with incorrect statements of the researcher being a graduate student). Our researcher was a professional research assistant hired by the university and with supervision by university faculty. She had a bachelor's degree in psychology and previous corrections research experience. She was trained to follow the standardized instructions from the manuals of the self-report tests. If we were doing psychological interviews, we would expect to use someone with a different background; however, the field researcher was capable of giving self-report tests as well as developing rapport with participants. One reviewer criticized her ability to determine if test scores were accurate and truthful. Her assessment of responses was related to response style (e.g., did the participant circle the same response for all questions) not response substance. This was not a judgment about psychological response but rather a check of how the participant filled out the score sheet. Assessments were marked questionable so we could consider those responses overall when analyzing an individual's consistency across measures or to check that there was at least face validity of taking the task seriously.

Although we recognize potential shortcomings of information from self-report data, we instituted numerous checks to determine if the data we received were reliable, valid, and useable. We believe that these checks provided information about the quality of the data, and we disagree with those reviewers who discount the use of paper-and-pencil tests.

We wonder if the criticisms about paper-and-pencil assessments are more about the findings than the data technique. If data from any one time period for the AS groups were used as the sole basis of our findings (i.e., if we used the same design strategy as Haney and Grassian), results would have indicated similar findings to these other studies. That is, we found high elevations on measures of psychological distress at every point of testing. Thus our results with paper and pencil tests are similar to what others find; however, within the context of measuring early and multiple times as well as including comparison groups, our study just differs in how we attribute the cause of the elevated scores.

### *Study Groups*

There are several criticisms concerning the sample and comparison groups. To be able to make the strongest causal statements about the impact of solitary confinement on psychological well-being it would be ideal from a research standpoint to randomly assign first time inmates to AS or general prison confinement and then observe change in psychological well-being over time. In such a situation we would be able to differentiate the impact of prison confinement in general from AS, study people prior to any segregation experiences, and assess psychological well-being from a more accurate baseline place. However, from a human rights view it would not be ethical to submit inmates to the strictest conditions of confinement when there are no behavioral indications that warrant such confinement. Thus, in doing research in an applied setting, within the constraints of that setting, we had to make several decisions about whom to study, what was the best comparison group available to us, and whether we should disqualify some inmates from participating. We recognize that there are different ways to answer these questions and a different design might have yielded different results.

Participants were not assigned their confinement condition by the researchers but were placed following the CDOC's processes and policies. Because Colorado requires a hearing for classification to AS following behaviorally disruptive behavior, there is a delay between when the hearing is scheduled and placement in AS or back into general prison population occurs. During this delay, participants are in punitive segregation. Because of the similarity between punitive and administrative segregation, we instituted a pre-baseline assessment. It is inaccurate that subjects were in segregation for 3 months prior to their first testing as stated by Shalev and Lloyd. We recognized that this early placement in segregation conditions might impact the psychological well-being of participants, which is the reason we obtained measures as quickly as we could. Participants were confined in segregation for an average of 30 days prior to the first assessment.

Per policy, Colorado places violent, dangerous, and disruptive inmates in AS. In seeking an appropriate comparison group, we sought to identify offenders who were as similar as possible on acting out behaviors but without the AS classification designation. Advisory board members were involved in the consideration of what was the best comparison group. Many of the articles alleging psychologically

destructive effects of segregation specifically referred to long-term segregation, and the previous studies (e.g., Zinger, Wichman, & Andrews, 2001) revealed few ill effects of short-term (e.g., 60 days) segregation. Thus, the advisory board did not believe that the short confinement in segregation that comparison subjects might experience at the start of the study would be detrimental to the design. Selection of this group then meant that it would be possible for them to be placed in punitive or AS during the course of the study because of inability to control these types of behaviors. We could have chosen a group that was most likely to stay out of segregation; however, this might have led to other inequalities between groups and thus alternative explanations for any possible group differences. If these early segregations did have a negative impact, then we would expect that the general prison group would improve after release but the AS group may continue to worsen.

Additionally, as may be expected from a behaviorally disruptive group, our participants had a history of AS (and likely punitive segregation), with the group being placed in AS during the study period having a higher incidence of past AS experiences than comparison subjects (35% versus 23%). We do not know if these earlier segregation experiences contributed to elevated measures of psychological distress at the start of the study; however, they do not explain the lack of change, or improvement, over time. Scores, although elevated over normative data, were not at the ceiling or floor indicating that there was room for groups to change if AS continues to have severe consequences for psychological distress. In the report we use previous AS experiences to predict change over time (Tables 23 and 24); previous experience was not a significant predictor for any variables. For this publication *Corrections & Mental Health*, we reran all the analyses comparing mean change over time and included previous AS experience as a predictor. There was not a single outcome variable in which previous AS experience was statistically significant nor had an effect size that explained a meaningful amount of variance (>2%); thus we do not think that previous AS experiences explains the results.

In addition to segregation history, it is also claimed that the sample has an extensive history of incarceration. We are unclear what statistics the critics are using to make this claim as nearly 70% of participants were incarcerated for the first time and only 8% had 3 or more incarcerations. Comparisons between the eligible pool and study participants did not indicate that study participants had worse criminal histories than the eligible pool.

Several criticisms have been made that our results are inaccurate because we excluded inmates who did not have a reading level that would allow them to understand the consent process or the self-report measures. We must disagree with Shalev and Lloyd who claim that this exclusion resulted in an over-representation of participants with a high school diploma or equivalent qualification, at least for how it relates to those who were eligible for AS. All participants who received a hearing for AS placement were in the eligible pool; results comparing the eligible pool to the study sample (in Table 2 of the full report) as well as comparisons between those who participated and those who actively refused (Table 3) indicated that there were not statistically significant differences on percentage of persons with high school diploma or equivalent. Thus our study sample seems representative of the pool of inmates who had an AS hearing. Overall, less than 4% of the 1,083 eligible to participate were excluded because of language or reading inability. We acknowledged in our report that our results could not be generalized to inmates who were not represented in our sample (e.g., illiterate persons, women, juveniles); however, we do not agree that this discredits our entire study. If our study is to be discredited for this, then we suggest that many

studies would need to be disqualified as we know of few that used representative samples from the prison population (e.g., Grassian's findings would be discredited for only using participants who were suing the state).

#### *Other Methodology Considerations*

The Colorado AS prison we studied has program levels through which inmates must progress in order to release from AS to the general prison population. We agree with the reviewers that it would be useful to have information about programming levels and it was our initial plan to use prison logs to track the number of days at each level for each individual at each study period; however, the prison logs were incomplete and we were not able to get complete information on individuals' progression through programming levels. Because we did not determine that the logs would not provide this information until late in the data collection phase, we did not have another way to collect these data during the study period. Such information could provide guidance on whether increasing privileges may account for those who changed in different ways.

Some reviewers requested information about social context in which the participants lived. We agree that this provides useful information but it was not the main focus of our study. We did collect data on feelings of safety, attitudes towards AS, and fear levels in AS from a scale that we developed. We used these as predictors for understanding change over time (see Tables 23 and 24, pp. 76-77) although the variables were not consistent or strong predictors of change, there is evidence to suggest that these variables may be relevant to understanding how people change (e.g., people feeling safe in AS tended to have less negative change in anxiety and somatization but more negative change on hostility/anger control; people with more positive attitudes about AS had tended to have fewer declines in cognitive performance on the Trails task; people with more fear tended to have more negative declines on the SLUMS cognitive task). Thus further research in understanding the conditions under which inmates' psychological well-being deteriorates or improves seems warranted.

#### **Future Research**

As stated previously, we believe in the research process with each research study building from past research and new research being suggested. Our study was specifically concerned with the psychological impacts of administrative segregation at CSP over a one year period; we did not study adaptation, institutional safety, or release to the community. We did not believe that our study would answer all the questions about the psychological impacts of AS and we were careful to describe the limitations of our research and to suggest future research. There are several extensions we would like to consider with this study's data including examination of intraindividual change over time rather than focusing on mean differences between groups and time; examining case files to gather previous history on mental health crisis events and CSP programming phases; and exploring if there are individual differences which explain participants' differential patterns of change. We are in agreement with all the reviewers about the potential research questions that have yet to be addressed as well as replications needed in order to complete the suggested meta-analytical studies. We would encourage these reviewers and other researchers to develop or extend their own research agendas to address the important unanswered questions suggested by the reviewers. In particular we would welcome a discussion with anyone willing to

discuss a multiple state collaboration to research the features of incarceration, including segregation, that would lead to the most benefit for both the individuals subjected to these confinement conditions as well as the safety of inmates and correctional staff.

**Authors' note:** *Maureen L. O'Keefe, M.A., is Director of Research at the Colorado Department of Corrections in Colorado Springs, Colorado; Kelli J. Klebe, Ph.D., is Chair of the Department of Psychology at the University of Colorado – Colorado Springs; Jeffrey Metzner, M.D., is a clinical professor of psychiatry at the University of Colorado School of Medicine; Joel Dvoskin, Ph.D., A.B.P.P. (forensic), is an assistant clinical professor at the University of Arizona Medical School; and Jamie Fellner, J.D., is senior counsel for Human Rights Watch. Principal Investigator Maureen O'Keefe can be reached at (719) 226-4364 or (email) [Maureen.Keefe@doc.state.co.us](mailto:Maureen.Keefe@doc.state.co.us); co-Principal Investigator Kelli Klebe can be reached at (719) 255-4175 or (email) [kklebe@uccs.edu](mailto:kklebe@uccs.edu)*

## References

- Grassian, S. 1983. "Psychopathological Effects of Solitary Confinement." *American Journal of Psychiatry*, 140, 1450-1454.
- Haney, C. 2003. "Mental Health Issues in Long-term Solitary and "Supermax" Confinement." *Crime & Delinquency*, 49, 124-156.
- Klebe, K. J. 2010. "Long-term Solitary Confinement's Impact on Psychological Well-being: The Colorado Study." San Diego, CA: Symposium presented at the annual meeting of the American Psychological Association.
- O'Keefe, M., K.J. Klebe, A. Stucker, K. Sturm, and W. Leggett. 2010. *One Year Longitudinal Study of the Psychological Effects of Administrative Segregation*. Colorado Springs, CO: Colorado Department of Corrections.
- Zinger, I., C. Wichmann, and D.A. Andrews. 2001. "The Psychological Effects of 60 Days in Administrative Segregation." *Canadian Journal of Criminology*, 43, 47-88.