Behavior and Social Issues, 17, 86-110 (2008).

REPLY

NEW MYTHS AND HARSH REALITIES: REPLY TO PAUL ON THE IMPLICATIONS OF PAUL AND LENTZ (1977) FOR GENERALIZATION FROM TOKEN ECONOMIES TO UNCONTROLLED ENVIRONMENTS

Jerome C. Wakefield[1]
New York University

ABSTRACT: As part of a larger argument about why token economy treatment for schizophrenia was largely abandoned despite demonstrated behavioral gains, I (Wakefield, 2006) analyzed Paul and Lentz's (1977) classic study of social-learning treatment of schizophrenia, sometimes cited as the best in this field. I argued that it failed to demonstrate or even test generalization of gains to uncontrolled natural environments, a serious drawback in an age of deinstitutionalization. In his response, Paul (2006) rejects my contention and argues that there are three sources of data in the study that support generalization: gains were maintained during a no-treatment baseline at 4 years into the study, during an 18-month period following a change in aversive time out procedures for aggressive acts, and during the 18-month follow-up of patients released to community aftercare. In this reply, I examine Paul's counterarguments and argue that the evidence strongly supports my original contention that Paul and Lentz's study provides no support for generalization.
KEYWORDS: schizophrenia, behavioral treatment, token economy, generalization, philosophy of science, history of psychology

Gordon Paul has devoted a good deal of print over the past thirty years (see the bibliography in Paul, 2006) to cultivating a mythical status for Paul and Lentz's (1977) study comparing social-learning token-economy treatment of hospitalized chronic schizophrenics to usual hospital care and an alternative milieu form of psychosocial therapy. The study was published in a lengthy book; its main results have never withstood the test of serious peer review in a frontline journal as far as I know, and have never been systematically empirically followed up. Nonetheless, the study is impressive in some respects and certainly important, and has sometimes been cited by others as the best the empirical literature on token economy treatment of schizophrenia has to offer.

Even a minor weakness in a mythical figure is unacceptable, I guess. So, when I noted what seemed some manifest limitations of Paul and Lentz's study in an earlier

[1] Correspondence may be directed to: Jerome C. Wakefield, Silver School of Social Work, New York University, 1 Washington Square N., New York, NY 10003. Phone: 212-998-5934. Email: Jerome.Wakefield@nyu.edu.

article (Wakefield, 2006), even though I praised the study's legitimate contributions and my comments were not about a central claim of the study, Paul (2006) took vigorous exception. He vehemently rejected my central contention that the study's data do not demonstrate or even rigorously test whether such programs' beneficial effects are likely to generalize to natural community environments where control over reinforcers cannot be maintained. To underscore his disagreement, he refused me permission to reprint a central table from his book, and threw in so much invective that one imagines he intended it as an ironic caricature of bombastic and excessively nasty academic prose.

In this article, I reconsider Paul and Lentz's (1977) study in light of the arguments (and accusations) presented in Paul's response. I rely here, as I did in my earlier discussion, entirely on the data and discussion in Paul and Lentz's own report.

It is an understatement to say that Paul is harsh in his assessment of my comments on his study;[2] he even suggests that my account is distorted "to a degree suggesting harmful dysfunction" (p. 244). Rather than referring me to a mental health professional for treatment, Paul generously endeavors himself to minister a "cure." Certainly, if dismissive, insulting rhetoric constituted evidence (or was therapeutic!), Paul's commentary would vanquish my arguments (and cure me for life!). The severity and number of Paul's accusations of scholarly weaknesses in my article warranted a full reexamination of my and Paul's claims. The Editor has thus generously allowed me to reply at length to Paul's response.

Despite our heated dispute, as Paul suggests, there are likely many overarching points about which he and I agree—likely many more than the points I share with the authors of the articles (Wong, 2006a; Wyatt & Midkiff, 2006a) that were the target of my commentary. It would be interesting to explore such common ground. But, like Paul, I focus here on the issues that divide us about the interpretation of his study.

It should be stated at the outset that, although I offer harsh verdicts about Paul's counterarguments, some of the "harsh realities" of my title are mine; Paul is correct in some of his criticisms, as I note below. In tackling his lengthy study report, I got some points wrong or stated them out of context in a misleading way. However, none of these errors impacts on the logic or validity of my argument. Moreover, in the course of rejecting my comments, Paul overreaches and creates some "new myths" about his study.

---

[2] For example: "Wakefield's overall critique is typified by a lack of theoretical understanding of his presumed theoretical enemy and its evidence base, selective perception to a degree suggesting harmful dysfunction, out-of-context quotes, and arguments based on factually incorrect allegations—clearly so in his assertions regarding the procedures and evidence in the Paul and Lentz (1977) monograph" (p. 244); "Wakefield's critique descends to a level of misinformation, declarations based on factually incorrect allegations, out-of-context quotes, and convoluted conclusions that are unequalled in any published "scientific" document that I have seen in more than 40 years in the field" (p. 245); "Wakefield slips in a truth and some half-truths before descending into misinformation and declarations based on factually incorrect allegations. Quotes and clarifications are provided below to show how his selective reporting and typical failure to provide any appropriate contexts reveal his biases—even in the rare circumstance when his statements are not technically incorrect" (pp. 245-246). And that's just the beginning!

Context of the Debate

It will be helpful to place the discussion in context by explaining how I came to be in the unenviable position of debating Paul about whether his own massive study of token economy treatment of schizophrenia demonstrates generalization of gains to natural community environments. In a special section in a previous issue of this journal, Wong (2006a) and Wyatt and Midkiff (2006a) published target articles in which they critiqued biological approaches in psychiatry and argued that the eclipse of behavioral treatment of severe mental illness is due entirely to the politics of psychiatry and not to any scientific considerations, and consequently that a return to behavioral treatment methods is warranted. In my commentary (Wakefield, 2006), I disputed the "politics" thesis. I argued that, at least in the central case of token economy treatments for schizophrenia, there is a scientific consideration, namely, limited and uncertain generalization of gains made in controlled environments to natural environments, that could explain the eclipse of behavioral intervention in an age of deinsitutionalization.[3] To support my point, I reviewed some of the literature on token-economy treatment of schizophrenic inpatients and underscored the limited generalizability that had been found by behavioral researchers. I then considered at length the implications for generalization of Paul and Lentz's (1977) study, which some consider the best in this field. I argued that this study, while supporting the effectiveness of token economy treatment in influencing to some extent behavioral expression of schizophrenia in a controlled setting, revealed weaknesses in generalization and stability of changes, and failed to provide a solid scientific case for generalization of gains to natural environments where control does not exist.

What happened next was strongly reminiscent of that famous scene in Woody Allen's movie *Annie Hall*, when the character played by Allen is standing in line for a movie with his date and they are suffering through a pretentious discourse on Marshall McLuhan's philosophy by an individual standing nearby in line. In irritation, Allen magically produces Marshall McLuhan himself, who proceeds to explain his views and to tell the pretentious individual that he does not know what he is talking about. In like manner, Wong, rather than entering into the discussion I initiated of the Paul and Lentz study, persuaded the Editor to invite Paul himself to newly enter the fray, thus initiating a separate thread of debate in which my commentary became the target article for Paul's critique.[4] No one offered me the option of bringing in my own intellectual

[3] Wong (2000b) and Wyatt and Midkiff (2006b) subsequently replied to my overall commentary, and I have responded to them previously (Wakefield, 2007) and they have replied (Wong, 2007; Wyatt & Midkiff, 2007).

[4] Paul in his commentary considers my comments only on his study. Referring to the other literature I discussed, Paul says that "I will leave it to the paper's author to rebut Wakefield's claims in that section" (p. 245). However, none of the target authors seriously addressed any of the literature on the token-economy treatment of schizophrenia and limits to generalization that I cited. Indeed, the target authors acknowledged in their replies that no confirmatory literature

gladiator/expert to champion my claims, so I was left to face Paul alone. Nonetheless, this was a welcome development. I was grateful for Paul's willingness to attempt to clarify the thorny issues raised about his voluminous study. In any argument, one wants to address the strongest case on the other side, and surely Paul can provide that case when it comes to his study.

However, just as, regarding the *Annie Hall* scene, some viewers questioned whether McLuhan himself knew what he was talking about, it must be evidence and argument, not authority, that holds sway in scholarly debate. So, rather than yielding to this seeming "deus ex machina" resolution, I now subject Paul's response to examination.

The Importance of Generalization

Paul agrees with me that, as Glynn (1990) says, "Maintenance and generalization of treatment gains are, of course, critical tests of the utility of treatment interventions" (p. 401). Paul adds that this is the "generally accepted premise of most clinical investigators" (Paul, 2006, p. 244), and complains that I cited Glynn's statement as if it were an indictment rather than a generally accepted fact.

Odd as it may seem, when considered in context, it *was* an indictment! The issue at hand was whether the target authors had supported their claim that the eclipse of behavioral interventions for severe mental illness was sheerly due to politics or had some scientific basis. The target authors argued that biological psychiatry, which now dominates, is weaker in its evidential base than it seems, thus the eclipse was due to politics and we should return to a behavioral paradigm. This reasoning is plainly fallacious; whatever the weaknesses of biological claims, one can conclude that behavioral methods were politically rejected only if one shows that they did not suffer from their own inherent scientific weaknesses. I pointed out that, in an age of deinstitutionalization, the main scientific challenge for token economy treatment of schizophrenia was considered to be generalization, and this remained problematic. As to the "indictment," despite the fact that (as Paul agrees) generalization was recognized as central, the word *generalization* does not occur in the target articles arguing for a resurgence of such treatment; the target authors simply ignored the question of whether behavioral treatments had been shown scientifically to be effective in this way.

Paul dismisses my concern about generalization with the "everybody knows that" defense, and reacts with horror to my claim that his study does not establish generalization. Yet, he had little scientifically serious to say about generalization in the 528 double-column pages (as he informs us at one point) of his book, other than that some techniques were used to promote generalization in the aftercare environment. If everyone knew that generalization to natural environments was critical to assessment of token economy treatment of schizophrenics, then why does the word *generalization* not appear in Paul's index, and why is there not even one chapter out of 43 explicitly devoted

---

exists, and that the empirical study of these issues more or less came to a halt early on. So, Paul is on his own.

to this topic? Neither do any of Paul's many subsequent reviews of the literature seriously address this issue.

The Literature According to Paul

Why, then, did I address Paul's study so seriously? If there were an extensive ongoing empirical literature demonstrating token-economy effectiveness when it comes to generalization to natural environments, there would be no need to look so closely at Paul and Lentz's (1977) thirty-year-old study. However, no such literature exists. I think, judging from Paul's various reviews, that Paul and I agree on this. He dismisses the other studies I cite as "early case reports and small developmental studies from the 1960s and 1970s," and observes that a successful test of effectiveness (let alone generalization of effectiveness to natural environments) would require longer periods of treatment than those studies allowed.

There are a few more recent studies than the ones I cite, but they come to much the same conclusion. For example, Wilder, White, and Yu (2003) treated bizarre vocalizations using differential reinforcement of competing responses and concluded that "the effects of the treatment package did not generalize from therapist A to therapist B. That is, the participant did not begin using the competing response with therapist B until after she was trained to use it by therapist B. Other studies (Wong et al., 1993) have also reported that, without specific training, skills taught to individuals with schizophrenia do not often generalize" (p. 50). (Note that Wong's 1993 study, which he cites in his reply to my original comment, is cited by these behavioral researchers as an exemplar of the failure of generalization.) Since Paul says that longer periods of training are necessary, he apparently *agrees* that these other studies, none of which (old or more recent) went on for years like Paul's study, were unable to confirm generalizability.

Perhaps it is worth tracing Paul's own reading of the scientific literature as stated in his own publications. Paul (2000, p. 6) refers to his 1997 (Paul, Stuve & Menditto, 1997) review as having synopsized the evidence for the effectiveness of the social-learning program (SLP). However, the 1997 article refers the reader back to the earlier Paul and Menditto (1992) review as having "established the SLP's superior efficacy, effectiveness, and cost efficiency…. The established effectiveness and promise of the SLP continued to receive empirical support through 1991, when Paul and Menditto (1992) completed their review" (Paul, Stuve, & Menditto, 1997, p. 3). No rigorous post-1992 evidence is presented in the 1997 article.

So the evidence apparently lies in the 1992 review. In that review, there is a section on studies of inpatient care from the 1960s and 1970s. Essentially, Paul observes that these early studies suggested effectiveness but suffered from methodological flaws, concluding that the methodological limitations meant that studies "of this sort cannot provide solid cause-effect relationships for a given treatment program; the program, per se, has not been manipulated alone somewhere in the design" (p. 51). He observes that "the reviewers of this literature were, however, nearly unanimous in calling for improved research design, controls, and description/specification of relevant variables to gain

adequate knowledge of the comparative effectiveness and limitations of inpatient approaches" (p. 51). The rest of the section on 1970s evidence is devoted to a detailed account of the Paul and Lentz study and its features and strengths.

What of the research in the decades since? The 1992 review offers a section on "Additional Evidence in the 1976-1990 Literature." In it, Paul and Menditto conclude: "Unfortunately, our examination of the published literature during the 1976-1990 period (concluded in February 1991) failed to uncover the hoped-for improvement in the scientific quality of individual studies. …Further, the spurt of activity through the mid-1970s appears to have fizzled…and became a comparative dribble by the late 1980s. Most of the publications during this period report the results of ongoing inpatient programs rather than prospective experimental studies….None of the more recent studies could provide firm evidence of comparative efficacy…" (pp. 53-54). They mention three useful studies, all from the 1970s. Nowhere in their review do Paul and Menditto offer any empirical studies bearing directly on generalization. The implication is inescapable: according to Paul's reading of the literature, his study stands almost alone in providing an evidentially rigorous foundation for claiming that social learning programs are comparatively efficacious. So, Paul's assessment of the literature on token-economy treatment of schizophrenia is about as negative as mine, and by implication he agrees that if the Paul and Lentz study does not demonstrate generalization, there is no other reasonable place to look to find rigorous evidence.

Overview

In the remainder of this article, I address the heart of the issue between Paul and me, namely, the degree to which the Paul and Lentz (1977) study provides rigorous scientific evidence for generalization of the positive effects of token economy treatment of schizophrenia to natural, post-treatment environments and the stability of such gains under environmental variation. The positive effects of such treatment within a controlled environment are not at issue here. Although one study cannot settle such an issue, in an age of deinstitutionalization it would be of interest if the Paul and Lentz study did provide some persuasive evidence on generalization.

There are three periods during Paul and Lentz's (1977) 6-year study that Paul cites as providing evidence for generalization of gains. First, 4 years after the treatment program began, all treatment was temporarily suspended during a 4-week no-treatment baseline in preparation for the next phase of the study. Paul claims that patient functioning during this period offers "strong evidence" for generalization.

Second, as I discussed at length in my commentary, there was a one and a half year period (between the fifth and eighth six-month assessments at 2.5 to 4 years into the study) during which there were changes in time out procedures for aggressive and assaultive behavior on the ward. I noted decreasing levels of patient functioning despite continued treatment during this period, and argued that this suggested environmental sensitivity of gains and thus less likelihood of generalization. Paul responded that I used

the wrong measure and misrepresented and misinterpreted the study, and that in fact the data regarding this period support generalization.

Third, the study included a one and a half year follow-up period after patients had been released to facilities in the community. I claimed that for various methodological reasons, the follow-up does not provide a scientific test of generalization. Paul denied my claim and argued that the follow-up in fact supports generalization. I assess the evidence from each of these periods in turn in the following sections.

PAUL'S ARGUMENT 1: MAINTENANCE OF GAINS DURING A 4-WEEK RETURN-TO-BASELINE WITHDRAWAL OF TREATMENT FOUR YEARS INTO THE STUDY SUPPORTS GENERALIZATION

Early in his response, Paul (2006) says, "Later I will provide evidence contrary to Wakefield's completely unsupported declaration that the Paul and Lentz study 'fails to demonstrate (or even to test) generalizability'" (p. 248). The most direct evidence he offers concerns the 4-week baseline period 4 years into the study. The baseline argument is of particular interest because it is the only time during the study that treatment was stopped and the results uncontaminated by continuing intervention. Paul argues that maintenance of improvements during the baseline's withdrawal of treatment and return to standard care offers strong evidence of generalization to new environments:

> Direct evidence of the generalizability of previous gains to a different environment was obtained from ongoing objective assessments of all areas of functioning during the four weeks in which return-to-baseline procedures removed all components of the SLP (tokens, programmatic material and social reinforcement, TO, etc.). This was done explicitly to test generalized improvements from the "old program" before reinstituting the new programmatic procedures, during the last six months of operation (Chapter 28, 29, & 31). Except that drugs were held constant at 11% during the 2nd one, assaults were handled by traditional hospital means during both the 1st and 2nd baselines—restraints, tepid baths, physical separation, and instruction. Staff were told to be pleasant, but to emulate the "aid culture" during both baseline conditions. The "82% improved" figure for SLP residents during return-to-baseline, following 18 months of operation without effective means of dealing with assaults, is *strong* evidence of generalization. (Paul, 2000, p. 250; emphasis in the original)

When Paul says that the "82% improved" figure during the return-to-baseline is strong evidence of generalization, the figure he cites does *not* refer to what happened to the patients from the beginning to the end of the baseline period (i.e., it does *not* indicate that, during the baseline period, patients improved). Rather, the "82% improved" figure refers to the percentage of patients that at the end of the 4-week baseline period remained improved *when judged relative to their status prior to the original introduction of the program four years before*. It is no surprise that during a brief 4-week baseline period in the same inpatient environment in which treatment occurred, the patients did not fully lose the gains of four years of treatment. To assess implications for generalization of treatment effects, one wants to know whether, during the baseline period itself, there

were any trends in the data that might reveal whether the new behavior had become functionally autonomous of the token economy and linked to natural reinforcers.

Notably, Paul (2006) remains completely silent on what actually happened to the SLP patients during the baseline period. The data are startlingly revealing; the social-learning group suffered statistically significant losses in almost every area of functioning during the four-week baseline period. Nor is this decline a matter of regression from the height of their achievement. As will be discussed in the next section (and is alluded to by Paul), the patients had already descended over the last year and a half from the heights of their functioning to a significantly lower level (but not as low as before treatment onset). During the return-to-baseline period, they decreased yet more from this lower point. In 4 weeks, not all was lost from 4 years of treatment, but the trend was clear.

What's more, these losses in the SLP group generally were not matched by similar losses in the milieu group. Thus, the rapid erosion of gains upon treatment withdrawal – that is, failure of generalization – was specific to the SLP learning.

Deteriorating functioning included increases in maladaptive behavior and decreases in adaptive behavior. Regarding maladaptive behavior, "social-learning residents showed an increase in clinically inappropriate behavior upon the return to baseline…. All three classes of concurrent, clinically inappropriate behavior (figure 29.2) increased at baseline 2" (Paul & Lentz, 1977, pp. 311, 312). In particular, of Cognitive Distortion (consisting of bizarre verbal and facial behaviors indicating thought disorder, such as delusions, hallucinations, incoherent speech, smiling without a stimulus), they say: "Cognitive Distortion showed a relatively greater increase for social-learning residents upon the return to baseline (figure 29.2)…, reflecting a significantly greater increase for social-learning residents than for milieu residents when the programmatic procedures were discontinued" (p. 312).

Moreover: "All three components of adaptive Clinical Frequencies (figure 29.4) [i.e., interpersonal skills, instrumental role functioning, and self-care] showed significant decreases for the social-learning group when the programmatic procedures were terminated during baseline…, while none of these changes were significant for the milieu group…. The decrease in the adaptive behavior of the social-learning group upon the return to baseline was significantly greater than parallel changes for the milieu group for both Self-Care and Interpersonal Skill…." (p. 316).

Paul inflates the fact that SLP patients did not in 4 weeks abruptly lose all gains and fall below the level they started at 4 years before into an argument for generalization. But significant decrements in functioning occurred across behaviors within 4 weeks especially in the SLP group, within the same physical environment between two periods of intensive treatment. Such decrements support the instability of gains at treatment termination specifically among SLP patients.

PAUL'S ARGUMENT 2: MAINTENANCE OF GAINS DURING THE PERIOD BETWEEN ASSESSMENTS 5 AND 8 (2.5 TO 4 YEARS INTO THE STUDY), WHEN TIME OUT PERIODS IN RESPONSE TO AGGRESSIVE BEHAVIOR WERE REDUCED IN LENGTH, SUGGESTS GENERALIZATION

In my earlier analysis, I argued that deterioration across behaviors in the SLP group between assessments 5 and 8 (2.5 to 4 years into the study) in response to one change in the the way the program responded to aggressive behavior suggested environmental sensitivity of gains and lack of transfer of behavioral support to natural-environment nonexperimental reinforcers, and augured ill for generalization. I relied on data reported by Paul and Lentz from the Inpatient Assessment Battery (IAB), administered at six-month intervals, to document the deterioration during this period, noting that this regression occurred despite continued treatment. According to Paul and Lentz, the deterioration was triggered by a change in an aversive procedure; after 2.5 years of treatment, the length of the time out period added to token penalties for aggressive behavior had, by state mandate, been reduced to 2 hours from 48 hours.[5]

Paul (2006) did not dispute my reasoning that if substantial deterioration occurred across behaviors because of one quite limited change in the SLP after 2.5 years despite continued treatment, this would suggest potential problems with generalization to uncontrolled environments. Rather, he responded that my reliance on the IAB led me to misinterpret the study's implications. He asserted that an alternative measure, Global Functioning (GF) ratings, collected on an ongoing basis, were superior in being more objective assessments and indicate a different conclusion:

> Wakefield completely misrepresents the degree, nature, and amount of loss for residents in the SLP following the statewide administrative reduction in time-out for assaults to 2 hours (216). By selectively reporting only the weakest measurement of all those involved, he claims the SLP showed "regression after years of treatment even as treatment continued"—completely disregarding three pages of analyses (pp. 377-379) that show how nearly everything he asserts as facts regarding absolute levels of functioning and comparative change are *actually artifacts.* The artifacts were due to clinical staff raters shifting their level of patient's ratings up or down on the basis of their own temporary emotional states. In an unbelievably prejudiced rejection of evidence, Wakefield (p. 217) dismisses as "a bit more positive in absolute terms in later assessment periods" the findings from the objective assessments of functioning that

[5] Paul [2006] correctly notes that I got some of the details wrong regarding the precise sequence of changes in the lengths of the time out periods over the course of the study, although this does not impact my argument. Also, Paul is arguably correct that my calling the longer time out periods and their associated aversive procedures "Draconian" was inappropriate, not because of a factual error as he claims, and not because they were not harsh or severe – they clearly were, and would not be allowed today – but because "Draconian" according to my dictionary involves a value judgment of "unjust harshness," and, as Paul argues, given the extremity of the patients' behavior and the fact that the study occurred at a time prior to increased concern about patients' rights, the severe nature of the time out process might be argued to have been justified in context.

> have been consistently documented to retain absolute levels of measurement over places and times (see Paul, 1987; Mariotto et. al., 2002).
>
> Most of Wakefield's above claims, in fact, are undercut by the objective data from the Time-Sample Behavioral Checklist (TSBC), gathered on stratified-hourly observations by trained independent observers, and from the Clinical Frequencies Recording System (CFRS), completed moment-to-moment by clinical staff (Chapters 24, 27, 42, 43). To be sure, the arbitrary reduction of the length of time-out and expulsion from 48 to 2 hours had a catastrophic impact on the Milieu Program and halted the trend of greater continuing improvements for the majority of residents in the SLP. However, contrary to Wakefield's assertions, this "natural experiment" actually provided evidence *for* the generalizability of previous gains to different environments for SLP residents.
>
> The group of SLP residents continued to show significant rates of improvement from their original levels of functioning throughout the 18-month period without effective consequences for assaults. Previous gains over initial levels were maintained in every area of objectively assessed functioning, *except* assaultive behavior. Because the uncontrolled increases in assaultiveness obviously resulted in a less therapeutic environment for everyone, maintenance of previous gains in a less hospitable environment is, arguably, evidence of generalizability. (p. 249)

New Myths and Spurious Accusations

Before addressing the substance of Paul's claims, I first respond to the accusations in this passage. Is Paul correct that my commentary "completely misrepresents" the results of the study by "selectively reporting" the weakest outcome measure? In fact, I simply followed and quoted extensively from Paul and Lentz's chapter 34, "Comparative Intramural Change of the Psychosocial and Hospital Groups," which contains their summary of the results of the 4.5-year intramural treatment using the IAB as the primary outcome measure. The only Figure and Table anywhere in the book summarizing the groups' changes for the entire intramural period appear here and use the IAB measure (the Figure is the one that Paul refused me permission to reproduce in my earlier article). No similar overall table or summary is presented for Global Functioning, although one could be constructed from the reports in various chapters. Other reviewers of the book (e.g., Liberman, 1980) also assumed that the IAB results were the study's primary outcome measure. But, it remains possible that GF scores reveal a different story, a possibility that I address below.

Is Paul correct that I proceeded by "completely ignoring" pp. 377-379 of the study report? Those pages contain a section titled "Some Minor Paradoxes on the IAB" that considers the divergences between IAB and GF scores. Far from ignoring the section, I summarized the relevant points as follows: "Paul and Lentz consider some divergences (which they call "minor paradoxes") between the IAB assessment battery measure and the ongoing day-to-day Global Functioning measures. In general, the Global Functioning

measures are a bit more positive in absolute terms in later assessment periods. But, they too show a substantial reduction in functioning (although continued significant improvement since initial baseline) of the social learning group" (p. 217).[6] Note that I explicitly acknowledged Paul's one repeated point, that GF scores indicated continued significant improvement from program onset. I did not elaborate this point because my argument regarding generalization did not depend on it. Rather, I was concerned with whether there had been significant and broad deterioration over the course of the 18-month interval in response to one specific change. The GF, like the IAB, did show significant losses across variables, a point that Paul and Lentz explicitly and repeatedly make themselves, and to which I return below.

So, there is no cogency to Paul's assertion that it is an "unbelievably prejudiced rejection of evidence" when I say that the GF ratings were "a bit more positive in absolute terms in later assessment periods" (note that Paul first says I completely ignore those pages, then argues that my summary of the pages was prejudiced). I appropriately acknowledged the piece of evidence on which his response rests—namely, the maintenance of absolute GF gains relative to the initial baseline of the study—but it was not an obstacle to my argument. Otherwise, I said what Paul and Lentz said, that the IAB and GF show parallel deterioration but the GF is slightly more positive.

Is Paul correct that the mentioned pages "show how nearly everything [Wakefield] asserts as facts regarding absolute levels of functioning and comparative change are *actually artifacts*…due to clinical staff raters shifting their level of patient's ratings up or down on the basis of their own temporary emotional states." Even if all of Paul and Lentz's speculations about the cause of GF/IAB divergences were true, the only "artifact," other than a slight shift in IAB absolute values, would be the IAB finding that functioning did not remain significantly above pre-treatment levels at assessments 7 and 8; as noted, the GF indicates that there was always a significant improvement from pretreatment. Other than this, the two measures offer quite similar portrayals of the patients' status over time.[7]

Although the question of absolute change from baseline is crucial with respect to assessing the success of the ongoing program in influencing behavior in a controlled environment, it is not crucial with respect to my argument regarding generalization. The continuation of treatment throughout this period could easily explain such continued

---

[6] Two other "paradoxes" mentioned in the section concern the directionality of changes prior to the fifth anniversary and parallels between the social-learning and milieu groups, neither of which are relevant to this dispute.

[7] As Paul and Lentz put the difference: "Continuous objective assessment of resident behavior found that the social-learning group showed significant improvement [relative to the pre-treatment baseline—JW] in overall Global Functioning at every anniversary assessment….In contrast, the improvement reflected in IAB assessments failed to achieve statistical significance on three occasions for the social-learning group. Thus, …the level of functioning on the IAB did not reflect the absolute level of improvement known to have been obtained from continuous objective assessments." (Paul & Lentz, 1977, p. 377) But, as noted, they also explicitly assert an overall parallelism of deterioration on the two measures during the 18-month period, documented below.

absolute gains. Reactivity to environmental change is still amply demonstrated by the broad deterioration that occurred in response to one program change (see below).

Is it true that I ignored the fact that Paul and Lentz (1977) "showed" that the GF/IAB divergence was due to rater reactivity on the IAB? In fact, Paul, and Lentz were engaging in post hoc speculation. They assumed without argument that the GF was the more valid rating ("the IAB did not reflect the absolute level of improvement known to have been obtained from continuous objective assessments" [p. 377]), without considering alternative hypotheses. Then, based on a post hoc analysis showing that judgments of staff about other variables correlated with levels of Intolerable Behaviors, they speculated that IAB scores were lower than GF scores due to rater reactivity.[8] Paul now confuses this speculation with scientifically established fact. Paul and Lentz clearly recognized the limits of their discussion at the time and included appropriate qualifiers ("suggest," "might underlie," "might be found," "seems probable") and say that it "appears" that reactivity may be suppressing IAB scores.[9] Paul now asserts those pages show rater reactivity, and even italicizes that the IAB levels were "*actually artifacts,*" as if italics can transform a speculation into a scientific fact.

---

[8] Paul and Lentz argued for this interpretation in the original report as follows:

> These analyses suggest that the incidence of Intolerable Behavior might underlie the differences in absolute level of IAB Functioning…compared to…objective assessments. [T]he basis for the paradoxical relationships and parallelism on IAB assessments might be found in reactivity of measurement….[T]he time-limited interview—by its very nature—is more subject to error arising from temporary disturbances of residents than are the continuous observational assessments that cover more representative periods of time and situations. Rating scales, on the other hand, are more subject to staff reactivity. Clinical Frequency data were also recorded by the same staff who provided ratings, but little opportunity exists for judgmental biases to enter since the presence or absence of discrete behaviors are recorded immediately upon their occurrence. In contrast, the rating scales require intermediate judgments of degree (rather than simple presence or absence) and recollection of broader classes of behavior over a period of several days—all of which allow for more reactive effects to enter from the staff who make the ratings. Thus, it seems probable that staff perceptions (and resident ratings) were somewhat negatively biased when staff were exposed to increases in physical assaults or to conditions that—based upon past experience—they anticipated might lead to increases in physical assaults. Such perceptual influences could result in residents' being accurately and reliably ordered relative to one another but with reactive ratings that shifted a few items by only one point, resulting in absolute level relationships similar to those obtained. (p. 379).

This series of speculations is the extent of the justification for the "artifact" claim.

[9] "[T]he absolute level of functioning on IAB assessments appears to have been suppressed for the psychosocial programs by reactivity to the incidence of dangerous and aggressive acts" (p. 379); "However, the incidence of dangerous and aggressive acts appeared to have reactive effects on measurement, which reduced sensitivity to other improvements and suppressed the absolute level of functioning on IAB assessments…" (p. 380).

Finally, Paul notes that "[Wakefield] claims the SLP group showed 'regression after years of treatment even as treatment continued,'" and he says this is an IAB artifact. He presumably means that the SLP did not regress below pretreatment levels. But, in the sense of "regression" relevant to my argument, there was indeed significant loss of GF over the 18 months and thus "regression after years of treatment even as treatment continued," as Paul and Lentz amply acknowledge in the original report (see below).

Harsh Realities: GF Ratings Indicate Patient Deterioration Between Assessments 5 Through 8

I now turn to the substance of Paul's counterargument in the passage quoted earlier. All of the above said, it remains possible that, as Paul asserts, Global Functioning measures support a different conclusion than mine. This claim is worth considering on its merits, and I do so in this section essentially by redoing my analysis based solely on GF measures. As before, I rely entirely on Paul and Lentz's (1977) report.

Paul's response relies on one specific divergence between the GF and IAB results reported in the study, namely, that GF retains impressive gains and never deteriorates to the pretreatment level during the 18-month period after reduction in length of time outs. Reading Paul's rebuttal, it might seem that he also asserts that use of the GF measures reveals that during the year and a half period, either further progress was made or SLP patients at least maintained the improvements that they had achieved up to that point. One would certainly never suspect from anything Paul says that the patients deteriorated significantly on GF during the 18-month period. He says things like "The group of SLP residents continued to show significant rates of improvement from their original levels of functioning throughout the 18-month period"; "[T]he arbitrary reduction of the length of time-out…halted the trend of greater continuing improvements for the majority of residents in the SLP"; and "Previous gains over initial levels were maintained in every area of objectively assessed functioning, *except* assaultive behavior." But, like Bill Clinton's "I never had sex with that woman," these statements are at best misleading; they all just mean that at four years of treatment there remained a significant difference from pretreatment.

Paul and Lentz's (1977) book tells a clearer story. Far from showing further improvement or a mere halt in progress during this period, GF significantly deteriorates across variables. Despite Paul's denial, the GF data clearly confirm my assertion that the SLP group showed "regression after years of treatment even as treatment continued."

The chapter concerned with this 18-month period of the study examines the GF ratings, and I follow its presentation. The title of the chapter, "The Next Year and a Half on the Psychosocial Units: Decline of Effectiveness and Search for a Cause" (p. 244, ch.23), already acknowledges the fact that Paul fails to mention in his response, that there was deterioration. In their book, Paul and Lentz make no bones about the occurrence of such deterioration, whether measured by GF or IAB. For example: "[T]he most notable feature of the current period was the relative loss of effectiveness of both programs that occurred after the change in consequences for assaultive behavior" (p. 264); "The

directionality of change between IAB Functioning and objectively assessed Global Functioning was uniform from the fifth anniversary assessment through the end of the intramural period; both psychosocial programs showed significant decreases over the year and a half" (p. 377); "The negative effects of reducing the length of time for expulsion and time out for assaultive behavior just before the fifth anniversary assessment in psychosocial programs had been evident over the following year and a half in continuous objective assessments. IAB Functioning similarly reflected significant declines over this period" (Paul & Lentz, 1977, p. 380).

The SLP group was continuously treated on the inpatient ward for two-and-a-half years, with great gains. Except for a change in severity of punishment of assaultive behavior, treatment continued as before for another 18 months. So, how widespread among the SLP patients were the significant losses in GF that resulted? Table 24.1 (Paul & Lentz, 1977, p. 262) informs us that, for the original SLP group (N=28), the rates of statistically significant change in overall GF were as follows: Worse, 57.1%; No change, 28.6%; Improved, 14.3%. There is no way to argue that these results support generalization to different environments. Needless to say, natural environments—and these days even controlled environments—do not generally allow for use of reinforcers of the types or levels of punishment that existed prior to the time-out change.

The increase in aggressive behavior that resulted from shorter time outs was enough to more than wipe out all gains in this area since program initiation 2.5 years before: "[B]y the end of the period, both programs showed significant increases over levels of intolerable behavior existing even before the program introduction….The social-learning program similarly found the most extreme resident to have changed from a zero incidence of Intolerable Behavior during the two weeks before the change to 36 during the two-week period of the eighth anniversary assessment. Concurrently the number of social-learning residents committing at least one Intolerable Behavior during a two-week period increased from 25% to 79%" (p. 256). The potential implications for transfer to uncontrolled community environments are obvious.

More interesting from the perspective of likely stability and generalization of gains to uncontrolled natural environments is that, while substantial gains from pretreatment were retained in many areas as treatment continued, the change in severity of punishment of Intolerable Behavior still had a major impact on other areas of functioning. Some other negative behaviors increased and virtually all positive behaviors decreased: "[S]imilar losses in previous effectiveness occurred in both maladaptive and adaptive components of functioning…" (p. 261). This suggests the environmental sensitivity of the changes brought about by the program.

Regarding maladaptive behaviors, while Hostile-Belligerence and Cognitive Distortion did not change significantly, Schizophrenic Disorganization and overall inappropriate behavior increased significantly: "[T]hey both showed significant increases in Schizophrenic Disorganization [primarily bizarre motor behaviors such as rocking, repetitive movements, blank staring] from the beginning to the end of the period….[T]he average social-learning resident had increased from about 24% to nearly 30% by the end of the period….considerable reductions from the 61% occurrence of Schizophrenic

Disorganization before the program introduction…" (p. 253); "The parallel increase for social-learning residents [in Total Inappropriate Clinical Frequencies] reflects a change from less than 3% occurrence at the beginning of the period to over 11% at the end—still less than the 19% average incidence before the program introduction…[the social-learning program] also showed increases in the incidence of performance of clinically inappropriate behavior…from 33% to 45% for social-learning residents…well below the 90% incidence rate that had occurred before the program introduction" (p. 250).

More surprising are the significant "losses in all four classes of adaptive behavior" that had been reinforced for 2.5 years and continued to be reinforced during the period of their decrease. For Total Appropriate Clinical Frequencies, there was a "steady reduction…over the entire period"; "The decrease reflects an average change for social-learning residents from over 61% terminal-level performance of social and instrumental behaviors during the fifth anniversary assessment to nearly 50% during the eighth assessment—still well over double the initial level before the program introduction" (p. 257). The specific categories included:

"Social-learning residents had shown a marked improvement in Self-Care [e.g., personal appearance, meal behavior, bathing, maintenance of personal living area] by the beginning of the period—to nearly 50% normal performance. By the end of the period, they had decreased to nearly 39% normal performance of Self-Care—a significant loss, but still three times their initial rate…" (p. 261);

"Instrumental Role Performance [e.g., "on task" in classes and job training and "on time" at scheduled work, activities, and meetings] had been initially less deficient…[T]he social-learning program had maintained its earlier improvements at a rate of over 81% normal performance during the fifth anniversary assessment. Both programs also showed losses in this class of adaptive behavior over the current year and a half…" (p.261);.

"Interpersonal Skills also showed a significant decline over both programs from the beginning to the end of the period…On an absolute level, the behaviors entering the Interpersonal Skills Index—relative frequencies of interpersonal interaction and communication skills—had originally been the most deficient of all areas of functioning [about 10% of opportunities]….While social-learning residents had shown dramatic improvements to nearly 48% normal performance of Interpersonal Skills by the fifth anniversary assessment, they also showed losses over the current period—to less than 40% normal performance on the average" (p. 260-261).

Finally, Concurrent Appropriate Behavior, encompassing twenty-seven classes of normal behavior, including facial expressions (e.g., smiling with apparent stimulus), positions (e.g., walking, sitting), and elective activities (e.g., grooming, writing), also decreased significantly during this period.

In previously assaultive patients, the suddenness and degree of aggression retriggered by changes in punishment regimens, as well as the concomitant change in other behaviors, is striking. Paul and Lentz offer an example, Bobbi, who "showed responses fairly typical of those other social-learning residents who had demonstrated assaultive behaviors before the reduction in the length of time out" (p. 266):

> [A]t the fifth anniversary assessment…her clinically bizarre behaviors had been reduced to less than 18% of her initial levels, and her concurrent adaptive behaviors had nearly doubled. All areas of adaptive functioning had increased from the nearly complete deficits existing before entry to the program to performance of self-maintenance and social skills indistinguishable from normal ones on nearly 60% of opportunities. Bobbi had performed only one instance of Intolerable Behavior in the four weeks before the change in time out was imposed. Overall, as with many other social-learning residents, Bobbi had literally changed from functioning as a savage animal to a pleasant human being, with quite a sense of humor.
>
> [U]pon the introduction of the change…a gradual increase in bizarre motoric behaviors started almost immediately….Once she performed an assaultive act and experienced the shortened time out consequence, Bobbi rapidly accelerated to a level of nearly one Intolerable behavior per day, accompanied by rapid deterioration in the performance of all adaptive behaviors and increases in bizarre motoric behaviors—actively avoiding other residents on over 87% of waking hours….By the seventh anniversary assessment, Bobbi was performing nearly two intolerable acts per day… (p. 266)

One sees here that what from a natural environment perspective is a relatively minor perturbation in circumstances yields a devastating deterioration in the gains made over years of treatment within weeks. Perhaps it is understandable that this might occur to someone who had a tendency to assaultive behavior to begin with. Thus, it is useful to consider an example Paul and Lentz present of what happened to a social-learning patient typical of several others who had not previously engaged in assaultive behavior:

> Olivia J (social-learning) showed responses to the events of the next year and a half that were typical of those of several other social-learning residents who had never engaged in assaultive behavior…. [After some ups and downs] By the fifth anniversary assessment she was again performing self-maintenance and social skills at a normal level on nearly 63% of opportunities, and clinically maladaptive behaviors were occurring on less than 39% of observations. Throughout the entire period before the imposed change in time out, Olivia had not performed an aggressive Intolerable Behavior.
>
> Within two weeks of the reduction in the length of time out, Olivia had performed an Intolerable Behavior and gradually increased the rate to one incident every other week. Through the seventh anniversary assessment, Olivia varied between no instances of aggressive intolerable acts and biweekly incidents—largely as a result of the presence or absence of senior staff coverage of off-hours…. [H]er aggressive Intolerable Behavior increased to a new high of over ten incidents per week. Thus, for Olivia, and several other social-learning residents who had not previously engaged in assaults, the reduction of consequences for assaults, combined with the increase in assaultiveness from other residents, appears to have established conditions for observational learning of assaultive behavior…. *[T]he externally imposed change in a single procedure had dramatically reduced the effectiveness of the overall program.* (p. 267; emphasis added)

I don't know how these GF data could be interpreted as anything but cautionary tales that gains in these programs are environmentally sensitive and may be unpredictably unstable under specific environmental changes, which in natural environments are everyday occurrences. Moreover, the environmental sensitivity occurred even in gains not directly related to the nature of the environmental change. It makes little difference whether one uses the IAB or the GF as outcome measures to make this point.

My central point was that the reaction to the time out change revealed likely environmental sensitivity of gains, casting doubt on generalization to uncontrolled environments. Despite Paul's protests, the fact is that Paul and Lentz (1977) reached the very same conclusion regarding environmental sensitivity of gains: "The TSBC Appropriate Behavior Index *appears to be exceptionally sensitive to environmental and psychological influences*, with the primary mediating variable during the current period being the increases in assaultive behavior….There were reductions in both programs over the current period…" (p. 260; emphasis added); "The detrimental effects of the changes in consequences for assaultive behavior were reflected in the losses of previous gains for adaptive behavior in both programs. Concurrent appropriate behavior again showed *exceptional sensitivity to environmental and psychological influences*….The classes of adaptive-behavior assessing components of the resocialization and instrumental role targets of rehabilitiation also showed losses during the period" (p. 262; emphasis added).

Paul (2006) tries to minimize the results by saying that "the aversive aspects of the SLP play an important, but limited role in the total program, which is a complex interactive package that is overwhelmingly devoted to positive, constructive actions" (p. 248). This is precisely what makes the response to the change so troubling and so suggestive of instability of gains under environmental changes. It is only one small component of a program largely devoted to systematic positive reinforcement of desirable behaviors, yet after 2.5 years of treatment a change in just the severity of the aversive component caused across-the-board significant reductions in functioning even as the program of reinforcement continued. The inescapable conclusion is that, although the aversive aspect is a limited part of the program in the literal sense of the proportion of interventions, it played a central role in facilitating the positive program. Paul and Lentz recognize this in their own analysis, which attributes the decline in all the positively reinforced adaptive behaviors to the less punitive time outs. The "natural experiment" established this point quite elegantly.

There is a final point regarding the environmental sensitivity of gains: Paul complains that I do not mention that adequacy of staffing was a major variable in control of assaultive behavior ("Wakefield fails to mention that, when staffing is adequate, overcorrection/restitution procedures are 100% effective in eliminating assaults that TO/token-fine procedures fail to control (p. 462)" [pp. 248-249]). Indeed, other than the primary factor of the change in severity of time out, changes in staffing is the variable most frequently mentioned in Paul and Lentz (1977) as influencing Intolerable Behavior and associated levels of functioning. Moreover, after the 18-month period discussed here was over, there was an upswing in patient functioning in the last six months of the study that coincided with having a full complement of study staff for the first time in the entire

study. However, the external controls brought about by adequacy of staffing (e.g., male senior staff on the ward at night) is precisely the kind of environmental variable that does not correspond to any stable feature of the natural environment, and suggests likely lack of solidity of gains under environmental variability.

PAUL'S ARGUMENT 3: MAINTENANCE OF GAINS DURING THE FOLLOW-UP PERIOD SUPPORTS GENERALIZATION

The third place that one might seek evidence regarding generalization in the Paul and Lentz (1977) study is in patient response during the 18-month community aftercare follow-up period. During this period, the study staff provided social-learning consultation to aftercare staff dealing with study patients who reached the threshold of functioning that allowed their release from intramural to aftercare facilities. In my commentary, I looked closely at the follow-up period and concluded that, although it took place in the community, it could not provide evidence regarding generalization to natural community environments. It was, as Paul and Lentz's chapter heading indicates, a "social-learning follow-up." The point was not to phase out token-economy treatment and see if gains were sustained in uncontrolled environments. Rather, the point was to transfer use of SLP techniques to the new environment by gradually phased-out consulting with aftercare staff. The community care facilities to which study patients were released were protected environments in which patient reward structures were still somewhat controllable; except for a few patients released to independent community living,[10] "discharges for all programs were to community board and room facilities with, at best, sheltered employment" (Paul, Stuve, & Menditto, 1997, p. 16).[11] Indeed, Paul himself seems to

[10] Only 11% of the SLP group were released to the community for independent living, not significantly different from the 7% so released from the milieu group, so no conclusions about what would have happened under such conditions to the SLP group overall can be drawn from those few releases. Interestingly, Paul, Stuve, & Menditto (1997, p. 16) assert that 25% of the SLP patients had functioning scores that had improved to "levels indistinguishable from 'normal'" by the time of release, but do not explain why nonetheless only 11% qualified for independent-living community placement, making one wonder what "indistinguishable from normal" means.

[11] Paul takes me to task for arguing that methodological problems made the interpretation of the SLP's generalization of gains to the community aftercare program difficult to interpret: "Wakefield's faulty assertions regarding supposed "methodological disasters" during follow-ups…simply shows a failure to read or understand the facts in Chapters 35-43. The *only* methodological problem with follow-ups was that we were unable to carry out a planned comparison of *two different modes* of declining-contact aftercare based on social-learning principles (case vs. program consultation)" (p. 247).

I pointed to two serious methodological problems. First, the retention in the community of extremely deteriorated patients (see below) makes the use of rehospitalization as an outcome variable more or less meaningless. Second, the administration of drugs to most aftercare patients, many of whom, especially among the social-learning group, had not been on medication at release, placed any causal inference about generalizability of SLP gains in doubt. Paul and Lentz (1977) themselves carefully report rates of medication over time during the intramural part of the study as

imply the essential role of the continued provision of treatment in sustaining SLP patients' gains when he notes: "Temporary declines in functioning for dischargees from inpatient programs (*only* those who experienced a reduction in social stimulation and an indiscriminant increase in prescribed drugs) were reversed once social-learning aftercare was begun as part of the planned generalization training to community environments and natural support systems" (2006, p. 247).

The results regarding overall patient response were impressive and argue for the usefulness of such consultation programs. On average, patient functioning did not deteriorate significantly over the course of the 18-month follow-up: "[G]roups who achieved significant release showed no significant loss from prerelease to the end of the follow-up period and still showed significant improvement from pretreament levels" (Paul & Lentz, 1977, p. 407). The immediate response to release was a reduction in functioning: "The mean decline in functioning of released residents from both psychosocial programs over the first six months in the community (-4.61) was significant (p's < 0.01), and both showed significantly greater losses than the loss of hospital releases over the first six months out…" (p. 406).[12] However, by the end of the 18-month follow-up period average losses were no longer significant.

---

a possible confounding variable. Regarding the follow-up, they suggest that the use of medication may have interfered with what would have been a better post-release transfer of learning from the token economy group (p. 441). Paul's vehemence notwithstanding, and irrespective of whether one suspects that the medication hurt or helped in community retention, the introduction of "indiscriminant" (Paul and Lentz's characterization) medication at the time of community release is clearly a serious confound that makes straightforward interpretation of the follow-up data with regard to generalization of SLP gains challenging. Paul's assertion that "The *only* methodological problem…was that we were unable to carry out a planned comparison" (p. 247) is incorrect.

Unhappily, I also asserted that Paul and Lentz had acknowledged this point: "Paul and Lentz themselves note in considering generalizability, '[T]he indiscriminant use of psychotropic drugs, severely limits any other conclusions' (p. 410)" (p. 215). However, Paul is entirely right when he complains that this quote in context "actually refers to our inability to draw further conclusions about the two modes of declining contact psychosocial aftercare based on social-learning principles rather than having *anything* to do with generalizability" (Paul, 2006, p. 247). But, despite my inadvertently misleading citation of Paul's supposed acknowledgment of my point, the essential point itself stands; in fact, the widespread initiation of medication in aftercare does obviously muddy up the waters in assessing what caused the outcome.

[12] To explain the initial greater loss of functioning in psychosocial than in standard-care patients, Paul and Lentz hypothesized two factors: "[T]he change from nondrug to drug state may have interfered with prior learning to some degree for psychosocial releases. The relative changes in attention and activity may have served to support new behavior for the hospital releases but not for previously acquired behavior for the psychosocial releases" (p. 407). The drug explanation as a major cause of the deterioration is questionable. Drug use in the psychosocial groups did increase dramatically in aftercare whereas in the hospital group it continued as before; in the last six months of the intramural program, 100% of the hospital group but only about 11% of the social-learning group were receiving medication, and all of the latter but only 43% of the former at a low dosage, whereas the vast majority of all groups received high doses of medication in aftercare.

Based on nuances of the data, Paul and Lentz argue that the primary reason for the overall success was the provision of social-learning consultation to aftercare staff, and thus continued effective use of social-learning methods in aftercare facilities. This result supports the beneficial effect of the aftercare consultation program and the power of social-learning treatment to influence patient functioning. However, it has no obvious implications for generalization or transfer of reinforcement to natural reinforcers in the community environments because treatment continued in the new environment via community proxies.

Paul claims the following support for generalization: "Following declining-contact aftercare, objective assessments during the last week before discharge predicts in-community ratings of functioning at 6, 12, and 18 month follow-ups, with correlations in the ±.60s and ±.70s—evidence of generalization unprecedented in the institutional literature" (p. 250). These data suggest that how people rated relative to each other after intramural treatment resembled how they rated relative to each other after the aftercare-plus-consultation program. But treatment continued, so such correlations could represent continued levels of differential responsiveness to treatment. Such correlations do not address whether the absolute level of functioning would be maintained in non-treatment environments. Generalization is not primarily a correlational issue of this sort.

A central finding of the follow-up was that patients from all three groups did equally well in staying in the community once released, a fact examined further below. Thus, as Liberman (1980) and Glynn (1990) noted, whatever was responsible for keeping SLP patients in the community could not have been the differential effectiveness of SLP gains in generalizing to the changed conditions in the community setting. The evidence compels Paul (2006) to follow Paul and Lentz (1977) in also recognizing that outcome during aftercare cannot be explained by differences in intramural treatment. Paul thus attributes the aftercare outcome to social-learning aftercare consultation irrespective of previous type of treatment: "*Such generalization training to community environments and natural support systems does appear to be the key to community tenure, no matter how patients achieved the level of functioning that qualifies them for discharge"* (p. 248; italics in original). Paul and Lentz (1977) also recognized that the critical factor must originate in the aftercare environment:"[T]he active psychosocial aftercare consultation had beneficial effects regardless of prior treatment conditions" (p. 407); "[T]he major impact appears to have come from the introduction of active psychosocial aftercare consultation, no matter what the mode of delivery" (p. 410).

Yet, the same percentages of patients (32%) significantly deteriorated in all three groups during that first six months of community placement. If, as Paul and Lentz suggest, the initiation of medication interfered with previous learning that took place under non-drug conditions, one would have expected a marked disparity in the rate of deterioration among those who learned under drug-free conditions and were being newly remedicated as compared to those who were continuing medication as usual and thus had already suffered whatever impact medication might have, yet no such disparity emerged.

The important thing to notice here is that, in attributing community retention to the aftercare social-learning consultation, Paul (2006) and Paul and Lentz (1977) reject any explanation based on the differential generalization of SLP gains. They thereby acknowledge that there is no evidence in the aftercare data of substantial differential generalization effects of social-learning gains from the intramural treatment. If, as all observers agree, the data indicate that the cause of the maintenance of gains in aftercare occurred during aftercare itself, then there is no evidence of SLP generalization.

Unraveling the Mystery of a Perfect Outcome

By far the most remarkable finding about the follow-up is that virtually none of the released patients from the three treatment groups were rehospitalized over 18 months: "[O]f all significant releases from the original equated groups, only one resident failed at community functioning and required reinstitutionalization. That one resident was a male from the social-learning program who had remained in the community for over a year at the time he was arrested for attempting to cash a forged check…" (p. 412).

One could easily get confused on this crucial point due to Paul's tendency to "spin" his data. For example, in a list of evidence supporting the SLP approach, Paul (2006; citing Paul, 2000, p. 7) asserts that the SLP has the "Lowest rates of recidivism or rehospitalization (<3%)" (p. 250). The truth, accurately stated in the book, is that virtually everyone in all three groups who was released remained released, so the SLP had the same rehospitalization rate as the other programs, not the lowest rate: "The success of our declining contact aftercare approach in maintaining dischargees from all three groups with rehospitalization rates of less than 3% over periods from 18 months to more than 5 years is strong evidence for it's influence" (p. 248). In his list of "evidence," Paul apparently measures "rehospitalization" from onset of the 4.5 year inpatient treatment 6 years before, thus including in the "rehospitalized" patients those who were never released and inflating SLP rates because it had a higher rate of release. But in fact SLP patients did not have a lower rate of rehospitalization once released.

The nature of the variables that were operative in causing the almost-perfect outcome regarding rehospitalization is a central mystery of the study. Paul (2006) as well as Paul and Lentz (1977), we have seen, attribute the outcome to social-learning aftercare consultation. But can the SLP consultation really account for a perfect retention rate? I think not, and I turn to the challenge of unraveling the mystery of this perfect outcome.

The clue that there is another part of the story beyond the effectiveness of the SLP consultation program is revealed in a startling finding that has been largely ignored: By the end of the follow-up period, among the approximately one-third of patients across groups who deteriorated significantly from the time of release, a substantial and similar proportion of all groups was functioning *below the levels that existed prior to intramural treatment onset*:

> Even though the groups who achieved significant release showed no significant loss from prerelease to the end of the follow-up period and still showed significant improvement from pretreatment levels, one aspect of the change over the follow-up

> period was particularly notable. Those who declined in functioning – an equal proportion for all subgroups -- had losses such that nearly 21% of all significant releases were functioning at a lower level in the extended-care facilities during the last follow-up than they were at the time of initial rejection for community placement six years earlier.... [T]he standards for acceptable functioning in extended-care facilities had lowered over the project period. (Paul & Lentz, 1977, p. 407)

This discovery eliminates the explanation that the SLP consultation worked so well to maintain prerelease functioning that no one fell below the rehospitalization threshold. These cases involved the loss of all 6 years of gains and more, and both Paul (2006) and Paul and Lentz (1977) emphasize how absolutely abysmal was that pretreatment functioning level.

In my earlier article, (Wakefield, 2006), I suggested that the uniform community retention was likely due to widespread medication. Drugs were given to almost everyone, so it seemed a reasonable guess that the drugs allowed the aftercare facilities to retain them. Paul spent a paragraph attacking this claim, and he was correct to the extent that I went beyond the evidence yet asserted the claim as likely.

However, Paul argues not just that my "drug" thesis is undemonstrated but that the evidence shows it is wrong. His counterarguments, though, are all non sequiturs that do not directly bear on my claim. They are arguments either that earlier medication did not by itself improve patients or that it slowed down the rate of learning of new behaviors in the SLP group. But my speculation was that in a new and stressful environment, medication might have slowed down the increase in negative behaviors or the unlearning of adaptive behaviors already learned in the SLP or otherwise allowed retention in the community.

Whether or not medication played a role in community retention, it cannot be the whole story. Some patients were taking drugs before program onset six years before, when they were rejected for aftercare placement despite functioning at levels like those during follow-up. Thus, the "drug" explanation begs the central question: Why were substantial numbers of patients with functioning well below the earlier threshold for community placement nonetheless retained in aftercare to the end of the follow-up period?

It must remain speculative, but a possible clue to what further may have been going on occurs in a passing comment in Paul and Lentz (1977). They report that, at the beginning of the sixth six-month period of the intramural study (2.5 years into the study), certain political pressures came into greater prominence: "By the beginning of this period, the problem of the "revolving-door' mental patient had become of major concern in the state, as well as nationally" (p. 243). It seems possible that this political factor, operating in the minds of aftercare staff as they made their decisions about patient disposition in this high-profile demonstration research project (Paul and Lentz report newspaper articles and a much-publicized investigation of their institution by a state congressman) could explain the surprising results. Simply put, there were perceived political pressures to keep patients in the community, and perhaps aftercare staff implicitly responded to these pressures.

If aftercare staff were determined to keep patients in the community, how would that work? One obvious way might be a tendency to ignore the SLP's minimal-drug philosophy and use drugs indiscriminately in an attempt to retain everyone, keep problematic patients from being disruptive, and keep others from being disrupted (such use of drugs might be usual practice anyway); a second way might be to relax or even abandon thresholds for rehospitalization, so that all patients, no matter how much their behavior deteriorated, would be kept in the community. Indiscriminate use of medication and avoidance of rehospitalization irrespective of behavioral deterioration is what in fact did happen; all else is speculation. But this would certainly be the way to create dramatic although spurious community retention statistics that showed progress in the struggle against revolving-door hospitalization. There are few other hypotheses, other than sheer inertia, that would plausibly explain the striking anomaly that 21% of the subjects were functioning at the end of follow-up at a level lower than they had been when treatment began and therefore well below the threshold for community placement. In my view, pending a different explanation, this possibility casts the meaningfulness of Paul and Lentz's (1977) much-touted community retention statistics into doubt.

I raised the issue of declining criteria for community tenure in my commentary, and Paul responded as follows:

> Wakefield's…questioning "why the benchmark for release was not lower,"… simply shows a failure to read or understand the facts in Chapters 35-43….Community personnel, themselves, established the "benchmark" of functioning acceptable for community placements from all three inpatient groups; no "lower levels" were allowed! (p. 247)

Paul's last point that "no lower levels were allowed" is manifestly untrue and diverges from Paul and Lentz (1977): "[T]he standards for acceptable functioning in extended-care facilities had lowered over the project period" (p. 407). Paul's response that the community personnel themselves established the criteria for community placement and that releases at a lowered level were not allowed is a non sequitur; I was not asking who set the release benchmarks or whether releases to the community were allowed at some lowered level. Rather, I was observing what Paul and Lentz also observed, that, once released, patients were allowed to stay in the community even when functioning decreased to a level lower than that which had been earlier required for community placement.

The question I raised was thus entirely ignored by Paul. Yet, the question has important implications for this discussion. A major claim to SLP success relative to other groups has been that more SLP patients—indeed, virtually all—reached the threshold for release. However, the magnitude of the difference in the groups' release rates depended on how high the threshold for release was set; if the release criterion had been lower, then more patients from other groups would also have qualified and been released, and there would have been less difference between the groups' release rates. The data imply that level of functioning was not crucial to retaining patients in the community once released. So the question arises: If the criterion for functioning used to justify release to the

community from the beginning of the study had been set at the lowest level of the patients subsequently successfully retained in the community during the follow-up, what percentages of the three groups would have been released to and retained in the community? I don't know the answer, but many patients in the non-SLP groups who did not in fact qualify for release might have been released and retained under the relaxed criteria.

Thus, the large group difference in release rates—a prime point in favor of the SLP—begins to look potentially somewhat arbitrary. It remains true that the SLP patients achieved higher levels of functioning, but it is not clear whether these higher levels were necessary for release and maintenance in the community. Paul's dismissal of the issue does not lessen the logic of the question and the doubt it casts on the scoreboard "success" statistics for the SLP regarding community release and retention. Once one realizes that all patients were retained in the community even if they lost all 6-year gains, Paul's assertions that "The success of our declining contact aftercare approach in maintaining dischargees from all three groups with rehospitalization rates of less than 3%... is strong evidence for its influence" and that "*Such generalization training to community environments and natural support systems does appear to be the key to community tenure"* must be looked at with partial skepticism, at least as sole explanations of community retention rates. But as an explanation for the maintenance of the gains that were maintained—and thus as an *alternative* to an explanation in terms of generalization from intramural treatment—Paul is correct that aftercare SLP consultation appears to have been beneficial for some.

CONCLUSION

Paul cites three periods of his study—return to baseline, period with reduced time out, and follow-up—as providing evidence for generalization of SLP gains. This evidence evaporates upon examination and is inconsistent with the details of the Paul and Lentz (1977) report. Paul's relentless "spin" notwithstanding, Paul and Lentz's study provides no significant evidence regarding generalization of token economy treatment gains to natural environments, and in fact suggests problems with such generalization. The overall lack of such evidence, and negative tendency of the evidence that there is, provide a scientific reason why token economy treatment of schizophrenia was eclipsed in an age of deinstitutionalization.

REFERENCES

Glynn, S. M. (1990). Token economy approaches for psychiatric patients: Progress and pitfalls over 25 years. *Behavior Modification*, *14*, 383-407.

Liberman, R. P. (1980). A review of Paul and Lentz's psychological treatment for chronic mental patients: Milieu versus social-learning programs. *Journal of Applied Behavior Analysis*, *13*, 367-371.

Mariotto, M. J., Paul, G. L., & Licht, M. H. (2002). Assessment in inpatient and residential settings. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches (2nd ed.)* (pp. 466-490). New York: Oxford University Press.

Paul, G.L. (Ed.). (1987). *Observational assessment instrumentation for service and research–The Time-Sample Behavioral Checklist: Assessment in residential treatment settings, Part 2*. Champaign, IL: Research Press.

Paul, G.L. (2000). Evidence-based practices in inpatient and residential facilities. *The Clinical Psychologist, 53*, 3-11.

Paul, G. L. (2006). Myth and reality in Wakefield's assertions regarding Paul and Lentz (1977). *Behavior and Social Issues, 15,* 244-252.

Paul, G.L, & Lentz, R.J. (1977). *Psychosocial treatment of chronic mental patients: Milieu versus social-learning programs.* Cambridge, MA: Harvard University Press.

Paul, G. L., & Menditto, A. A. (1992). Effectiveness of inpatient treatment programs for mentally ill adults in public psychiatric facilities. *Applied & Preventive Psychology: Current Scientific Perspectives*, *1*, 41-63.

Paul, G. L., Stuve, P., & Menditto, A. A. (1997). Social-learning program (with token economy) for adult psychiatric patients. The Clinical Psychologist, 50, pp. 14-17.

Paul, G.L., Tobias, L.L., & Holly, B.L. (1972). Maintenance psychotropic drugs in the presence of active treatment programs: A "triple-blind" withdrawal study with long-term mental patients. *Archives of General Psychiatry, 27*, 106-115.

Wakefield, J.C. (2006). Is behaviorism becoming a pseudo-science?: Power versus scientific rationality in the eclipse of token economies by biological psychiatry in the treatment of schizophrenia. *Behavior and Social Issues*, 15, 202-221.

Wakefield, J. C. (2007). Is behaviorism becoming a pseudoscience?: Replies to Drs. Wyatt, Midkiff and Wong. *Behavior and Social Issues*, 16, 170-189.

Wilder, D. A., White, H, & Yu, M. L. (2003). Functional analysis and treatment of bizarre vocalizations exhibited by an adult with schizophrenia: A replication and extension. *Behavioral Interventions*, 18, 43-52.

Wong, S. E. (2006a). Behavior analysis of psychotic disorders: Scientific dead end or casualty of the mental health political economy? *Behavior and Social Issues, 15,* 152-177.

Wong, S. E. (2006b). Response to the commentaries. *Behavior and Social Issues, 15,* 232-243.

Wong, S. E. (2007) Scientific discovery, social change, and individual behavior change. *Behavior and Social issues*, 16, 190-196.

Wong, S. E., Martinez-Diaz, J. A., Massel, H. K., Edelstein, B. A., Wiegand, W., Bowen, L., & Liberman, R. P. (1993). Conversational skills training with schizophrenic inpatients: A study of generalization across settings and conversants. *Behavior Therapy, 24,* 285-304.

Wyatt, W. J., & Midkiff, D. (2006a). Biological psychiatry: A practice in search of a science. *Behavior and Social Issues, 15,* 132-151.

Wyatt, W. J., & Midkiff, D. (2006b). Six-to-one gets the job done: Comments on the reviews. *Behavior and Social Issues, 15,* 222-231.

Wyatt, W. J., & Midkiff, D. M. (2007). Psychiatry's thirty-five-year, non-empirical reach for biological explanations. *Behavior and Social Issues*, 16, 197-213.