Behavior and Social Issues, 15, 202-221 (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

Jerome C. Wakefield[1]
New York University

ABSTRACT: Wyatt, Midkiff and Wong argue that biological psychiatry's power, not its scientific merits, explain token economies' eclipse by biological treatments of schizophrenia. However, these critiques of biological psychiatry, while partly true, ignore offsetting strengths and achievements as well as plausibility arguments that schizophrenia is partly biological; behavioral theory offers no cogent alternative account of etiology. Moreover, token-economy research failed to establish generalizability of changes to post-ward environments. Even Paul and Lentz's (1977) definitive research on token-economy treatment of schizophrenia failed to show generalization of changes to community life, and in fact, due to an inadvertent "natural experiment," revealed the instability of behavioral changes even after years of treatment. To preserve their belief system, behaviorists seem in danger of turning behaviorism into a pseudoscience defended by ad hoc hypotheses.
KEYWORDS: behaviorism, behavioral treatment, schizophrenia, mental disorder, history of psychology, history of psychiatry, philosophy of science, harmful dysfunction, biological causation; pharmaceutical industry; psychotropic medications

After a period of provocative research on token economies in the treatment of schizophrenia, interest in this modality waned to the point where the field now barely exists. Wyatt, Midkiff, and Wong's contention is that biological approaches came to dominate over behavioral methods in the understanding and treatment of schizophrenia for largely unscientific, social reasons such as the political positioning of the medical model and the power of the pharmaceutical industry, rationalized by a flawed research base.

However, I see little in Wyatt, Midkiff and Wong's articles to support this contention, and I think by and large the contention is a myth, albeit a potentially powerful, community-sustaining myth. One must be suspicious of such special pleading. The claim that behaviorist approaches were set aside by the scientific community in favor of biological approaches not because the scientific limitations of behaviorism itself allowed the alternative to become dominant but rather for nonscientific reasons surely

[1] Correspondence should be directed to: Jerome C. Wakefield, 309 W. 104 St. #9C, New York, NY 10025. Phone: 212-932-9705. Email: Jerome.Wakefield@nyu.edu.

deserves a high burden of proof, lest every scientific debate be turned by the loser into an interminable sociological investigation.

Like any causal hypothesis, the hypothesis that the waning of token economy treatment for schizophrenia is due to unscientific forces needs to be carefully assessed with full recognition of the evidential complexity of such claims. Such a claim must be supported not merely by showing that biological psychiatry has weaknesses—so does every theory!—but by showing also that behavioral approaches do not have equal or greater weaknesses. Just because one can identify non-scientific forces at work, that does not mean that there was not an underlying scientific logic that allowed those forces to hold sway. Such extra-scientific interests and constraints *always* exist and *always* exploit a science-based shift, and thus are always identifiable by those seeking pseudoscientific preservation of a disconfirmed view. The question is whether such processes *replaced* science rather than *accompanied* a reasonably legitimate process of scientific progress.

I will argue that a balanced view of the achievements of biological psychiatry and the history of research on token economies suggests that there was sufficient rational warrant for an emphasis on biological methods over behavioral intervention. Thus, the shift noted by Wyatt, Midkiff, and Wong was at least in substantial part based on a scientifically legitimate reading of the evidence. This does not mean, of course, that the balance might not change again in the future with new research.

I first comment on Wyatt, Midkiff and Wong's critiques of biological psychiatry and argue that the biological psychiatric record, while suffering from many deficiencies, is not as weak as they claim. I then review some of the history of research on token economy treatment of schizophrenia, emphasizing the problem of generalization of changes to post-treatment environments. I pay particular attention to the much-cited study by Paul and Lentz (1977), arguing that this study does not show what it is often taken to show.

WYATT AND MIDKIFF'S CRITIQUE OF BIOLOGICAL PSYCHIATRY

I can address only a sample of Wyatt and Midkiff's arguments for the weakness of biological psychiatry's research base. Surely they are right to this extent: twin studies, brain scan studies, and other such approaches often have unacknowledged methodological weaknesses, do not show as much as biological psychiatrists claim, or have not had the degree of substantive payoff that was once imagined. Granted, as in many fields, we know much less than some biological researchers have sometimes claimed.

However, Wyatt and Midkiff's assessment of biological research is equally unbalanced in not acknowledging the power of the biological evidence. Wyatt and Midkiff's critique of the use of twin studies that are used to demonstrate genetic components of mental disorders is marred by the fact that the main target of this critique, studies using separation of monozygotic twins at birth, is important but of limited value in the overall scheme of things because such separations as well as mental disorder are both very rare, so the number of instances is limited. They ignore the main strategy in

such studies, namely, comparison of monozygotic versus dizygotic twin pairs and other comparisons of those with differing genetic and environmental overlap. It should be kept in mind that evidence for heritability and more generally for genetic and biological influences on schizophrenia is based on convergent evidence from a large number of sources, including not only twin studies and family studies but also, for example, higher rates of schizophrenia due to inbreeding, head trauma, and greater paternal age. Keller and Miller (in press) aptly summarize just one area of such evidence as follows:

> Using different methodologies, behavioral geneticists have consistently found that mental disorder heritability estimates range from about .2 to about .8, meaning that 20% to 80% of the differences between individuals in mental disorder liability are accounted for by differences in alleles between people. Without acknowledging genetic influences on mental disorder, only the most convoluted, post hic arguments could explain why (a) adopted children are consistently more similar to their biological than to their adoptive parents, (b) siblings and twins reared apart are about as similar as siblings and twins reared together, (c) similarity in extended families decreases monotonically as a function of genetic similarity, and (d) identical twins are consistently more similar than fraternal twins. (Keller & Miller, in press)

Wyatt and Midkiff's riposte is that monozygotic twins could be similar in ways that trigger similar environmental stimuli, thus similar learning, in precisely the (unspecified) areas that shape schizophrenic behavior. And, I suppose one might add, dizygotic twins could be just enough different that they trigger somewhat differing environments that produce lower but still significant levels of environmental concordance in ways that happen to be precisely related to schizophrenogenesis.

Surely Wyatt and Midkiff's critique is a clear example that "only the most convoluted, post hoc arguments" could provide alternatives to the standard conclusion. One can of course *assert* that environmental similarities occur in response to genetic similarities in just the needed pattern to explain the levels of symptom similarities in all these areas of research that behavioral geneticists routinely predict and find. But the authors offer no evidence for this speculation, and there is no cogent reason based on behavioral theory to expect such regularities, whereas the genetic hypothesis successfully *predicted* these confirmed patterns of responses. For every theoretical claim, no matter how well confirmed, one can generate alternative possible explanations. Until there is evidence for them, the more plausible and non-ad hoc explanation that led to the prediction is to be preferred.

Turning to the "unbalanced neurotransmitter" theory of mental disorder, I agree with Wyatt and Midkiff (and I think there is broad agreement among biological psychiatrists) that this "humoral" approach has been oversold, and perhaps for the worst of reasons. But that does not mean that schizophrenia and major depression do not involve other biological problems. I also agree that environmental influences are sometimes ignored in favor of naively strong assumptions about genetic determination (a naivete that behavior

geneticists do not generally share), and that we must challenge such overreaching. But it is equally true that blind faith by functional analysts in the existence of unknown environmental determinants, including unknown reinforcers that somehow manage to maintain seemingly painful and heavily punished behavior, should equally be challenged.

Wyatt and Midkiff conclude that the recent "rise of biological causation theory … has come about as a direct result of the powerful influences of organized psychiatry and the pharmaceutical industry." In fact, biological causation theory and the power of the pharmaceutical companies came about partly as a result of the gross failure of rival paradigms (including psychoanalysis and behaviorism) to successfully explain or adequately treat severe mental disorder, combined with the impressive results of modern psychotropic medications, whatever their negatives. The fact is that it was apparently chlorpromazine, not token economies, that emptied the asylums; it was lithium that reduced an enormous suicide rate among bipolar individuals to below general population levels; it was anxiety medication that relieved masses of individuals from the aversive effects of our stressful lifestyles as well as from anxiety disorder symptoms; and it was recent hypnotics that enabled millions of people to immediately experience a beneficial change in their "sleep behavior." These positive apparent realities cannot be ignored in a balanced analysis of the history of scientific judgment, even if negatives also attach to each of them, and even if, as inevitably happens, subsequent scientific scrutiny reveals weaknesses in these initial findings.

Wyatt and Midkiff also conclude that, "it is time for a paradigm shift, away from extreme biological causation and toward an environmental causation model, one that recognizes that at least some disorders are biologically caused." The last clause is a welcome nod in the direction of reconciling behaviorist aspirations with the complexity of reality. But, except as a straw man, there is no "extreme biological causation" model held by serious theorists or researchers. The general view, and all the evidence (including the twin evidence, as Wyatt and Midkiff know), is that there is gene-environment interaction. It is mostly functional analysts who have tended to resist an integrative conclusion. And, it must be kept in mind that "environmental" does not necessarily mean "operant conditioning."

At a more subtle level, Wyatt and Midkiff seem to assume that a social explanation is more politically progressive than biological explanation. Certainly, environmental explanation does suggest environmental change as a strategy, and that can be progressive. However, biological explanations can potentially cut both ways. One might argue that in the current environment, more resources can accrue for the treatment of mental disorders through the use of a biological perspective (thus, NAMI's campaign to classify mental disorders as biological disorders, and various government agencies' identification of those mental disorders that are biological disorders) because they are seen as more politically neutral, objective medical conditions. In contrast, environmental explanations tend to be inherently political because they inevitably bring in contingencies involving social conditions. This is one reason NIMH jumped to embrace a more biological way of talking when the Reagan administration came in. It seems better for mental health professionals to focus on the scientific truth of the matter than to become politicians.

Behavioral definitions of "disorder" underscore the political dangers of an environmental approach. Such definitions generally scoff at medical models, thus implicitly ejecting themselves from the domain of the health professions and calling medical reimbursement into question (Wakefield, 1998, 1999b). As a replacement for an objective biological dysfunction as the criterion for what is considered a disorder (Wakefield, 1992, 1999), these definitions generally refer to social values or the undesirability of behavior. While I would argue that anyone who is suffering should be granted support in the form of mental health intervention, one must admit that it is indeed conceptually unclear why all the sorts of problems described in behavioral definitions of disorder inherently fit within the mandate of the mental health system. The behaviorist attack on the medical model seems at odds with the relentless striving by behaviorists along with other psychologists to be considered legitimate mental health professionals qualifying for medical reimbursement.

WONG'S CRITIQUE OF BIOLOGICAL PSYCHIATRY

I now turn to Wong's article. Again, I can only address a sampling of his arguments, although I consider selected issues in more detail in subsequent sections.

Wong begins with the statement: "Behavior analysis once offered a bright promise for advancing the understanding and treatment of severe mental disorders." I address treatment later. Regarding the understanding of psychosis, Wong's statement is without foundation. It remains of interest that operant procedures can *influence* the manifestation of psychotic behaviors. But never did the literature on conditioning or token economies with psychotic patients suggest any serious understanding of the etiology of psychosis, any more than shaping the behavior of developmentally disabled children or head trauma patients offers a cogent etiological explanation of these conditions. (Ayllon and colleagues' success in getting a psychotic woman under controlled institutional conditions and with prompting by staff to hold a broom when inappropriate so that it appeared to psychiatrists to be a symptom is not serious evidence of general psychotic symptom etiology, in my view.)

The token economy literature, by showing that even bizarre psychotic behaviors are influenced by learning to some extent, did offer a corrective to ideological claims of rigid biological determination of psychotic symptoms. But Wong would be the first to point out that the fact that chlorpromazine reduces psychotic symptoms does not mean that the patient is suffering from a lack of chlorpromazine. Abandoning parity of reasoning, Wong (1996) elsewhere embraces explicitly what is here implicit, namely, the inference from behavioral influence on symptoms to behavioral etiology of symptoms. For example: "Whether initially established by positive reinforcement, negative reinforcement, or a combination of both, bizarre behavior in adults with psychosis often evolves into a variform and durable response pattern" (p. 326). Usually, though, Wong tries to verbally remain more neutral, suggesting that he is concerned about maintenance or perhaps just change of symptoms. These are each very different hypotheses.

Wong is not alone in equivocating on what, exactly, he is claiming. This kind of equivocation is common in the behavioral literature. For example, consider Wilder et al. (2001). The authors first state in the abstract that they are examining "variables responsible for the maintenance of bizarre vocalizations" (p. 65), despite the fact that the notion that the factors they use to *change* the behavior are the very factors that were maintaining the behavior prior to the study is wholly speculative; as noted, the fact that a behavior is sensitive to reinforcement is not support for the claim that the behavior was entirely or even partially brought about or maintained by reinforcement. Then they say, "Although traditional accounts of these behaviors posit that they are symptoms of an underlying disorder, behavior analysts view these behaviors as a class of operants that are influenced by environmental contingencies" (p. 65), thus setting up a false dichotomy; behaviors caused by a disorder may be influenced by contingencies. There is also an evasive fuzziness regarding what it is for a scientist to "view something as" a certain kind of thing versus showing that the evidence supports the claim that it is that sort of thing.

After noting earlier studies in which vocalizations were influenced with differential reinforcement, Wilder et al. (2001) state: "These results suggest that these vocalizations in some individuals with schizophrenia may be maintained by, or at least are sensitive to, social consequences such as attention and escape" (p. 65). The latter disjunction covers a vast territory; being *responsible for maintaining* a behavior under natural conditions prior to treatment is utterly different than being *capable of influencing the behavior* during treatment. The rhetorically swift move from "may be maintained by" to "or at least are sensitive to" is in fact a scientifically major move between a theoretically strong etiological or maintenance *explanatory* claim to a much weaker claim about being able to have an effect on a symptom. This sort of equivocation on what one is claiming obscures the structure of the argument and makes it easy to avoid paying the falsificationist piper.

At any rate, how does Wong's inferred behavioral etiological account handle the obvious facts of psychotic behavior, in particular the resistance to extinction and the resilience and maintenance of psychotic symptoms even when the environment and associated contingencies change dramatically (e.g., family-of-origin versus institutional commitment versus deinstitutionalized homelessness or community placement)? The answer he offers elsewhere, provided without a shred of independent evidence, is that the symptoms must be very very well learned ("overlearned") due to unknown reinforcers:

> Bizarre responses, most notably psychotic speech, will at times resist contingency management procedures…or will spontaneously recover over time…or when training has ended…. These results have been interpreted as showing that clients' underlying belief systems have remained intact despite behavioral training. However, multiform and persistent bizarre verbalizations can be parsimoniously viewed as generalized responses with a long history of intermittent reinforcement. After being positively and negatively reinforced by different people in various situations over many years, bizarre verbalizations could be overlearned responses that resist contingencies administered in circumscribed therapy sessions over mere weeks or months. Furthermore, it is

> difficult to prevent other patients, lay persons, and medical professionals from continuing to reinforce these verbalizations in the usual manner. (1996, p. 326)

Wong's rationale raises the question: When does "overlearning" become an excuse for not accepting disconfirmation of behavioral principles (as in attempts to evade the strong evidence in favor of biological preparedness for specific learning)? If we take Wong's rationale seriously as a testable thesis, then it appears to be disconfirmed, as I will argue later in my discussion of the Paul and Lenz study. Reminiscent of the evasive Wilder et al. framing discussed above, Wong asserts that symptoms "can be parsimoniously viewed as" learned behavior; but the parsimony exists only if one does not try to reconcile the account with the network of overall scientific knowledge. Symptoms "can be parsimoniously viewed" as byproducts of unknown biological brain dysfunctions, as well; the point is not how symptoms can be viewed, but how they are most scientifically plausibly viewed when one considers all the evidence.

Turning to Wong's critique of biological psychiatry, the lack of parity is striking. For example, he criticizes the lack of reliability of psychiatric classification, but never even raises the question of the reliability of a functional analysis.

Wong asks: "Why does psychiatry represent schizophrenia and other mental disorders as being brain diseases without first obtaining definitive evidence?" His explanation, borrowed from the the critical psychiatry movement, is the following: "Medical sociologists and other critics of the new biopsychiatry point out that this perspective portrays mental disorders as somatic problems and therefore the appropriate domain of medical practice, and that it rationalizes psychiatry's hegemony over the other mental health professions….This ideological shift also gave psychiatry a rich and powerful ally: The pharmaceutical industry."

Actually, psychiatry was under siege in the 1960s and 1970s from the behaviorist critique, the antipsychiatric movement, attacks on reliability, Rosenhan's study, psychoanalytic dominance, internal theoretical fragmentation, and so on. The resurgence in its power occurred *because* medication was found to be a powerful treatment for the severely disturbed, and also because psychiatry at least attempted to systematically address the criticisms.

Wong is correct in his concerns, in my view, to the extent that there is a dangerous trend toward using medication reflexively and exclusively, when in fact the reduction of symptoms through medication should almost always be accompanied by psychological or behavioral intervention to build new skills and capacities. But the fact remains that behavioral, psychological, psychoanalytic, and social methods of change all had their day and, though touted by devotees (and though no doubt every group of devotees would argue it was politics, not evidence, that undid them), all failed to basically change the situation of the severely mentally disordered.

Anyway, there is a pretty clear answer to Wong's question of why psychotic behavior is considered a medical disorder. It has almost always been considered a medical and even physiological disorder, ever since at least Hippocrates and Aristotle

through to Kraepelin's biological thesis at the turn of the 20th century. The reason has nothing to do with the recent factors mentioned by Wong, or Wyatt and Midkiff.

The resurgence of the biological view is based on traditional prima facie inferences about biologically designed normal human capacities across expectable environments, plus persuasive contextual evidence (i.e., the kind we commonsensically use when judging that blindness or paralysis is a disorder even when we know nothing about the physiology of the eye or musculature), plus a failure of all suggested alternative theories (including various social, psychological, and behavioral theories) to reach a minimal threshold of plausibility. In particular, the relative independence of psychotic symptoms from environmental change and intervention suggest an internal cause, and the failure of biologically designed functioning suggests a biological dysfunction. The retention of similar symptoms from family to institutional to community environments, plus the overwhelming evidence of at least some degree of biological etiology (see the discussion of Wyatt and Midkiff's paper), poses a powerful prima facie challenge to environmental explanations. Is it really possible that the reinforcers in all these varied environments are so aligned for these particular people, despite the heavy toll in suffering and social stigma and disapproval that psychotic symptoms usually incur and that should lead to extinction over time, that the explanation for the maintenance of the symptoms should be sought in a functional analysis? Or is it more likely that one can do a functional analysis to figure out how to reduce these behaviors, but that it has little to do with the genesis of the behaviors? Anything is possible, but pending breakthrough evidence to the contrary, most observers of psychosis throughout history have been persuaded of the biological position. The idea that there are unsuspected, hidden reinforcers that explain this behavior is simply less plausible than that there are unsuspected, hidden biological or psychological dysfunctions of some internal mechanisms that lead to the behavior.

Such inferences to likely causal domains are not remotely "definitive evidence," as Wong correctly notes. But, to address the other part of Wong's question, the fact that schizophrenia is judged (tentatively and fallibly) to be a medical disorder without definitive evidence is just routine science, and has to be judged on plausibility grounds. There is almost never conclusive evidence for any scientific theory; we are always inferring fallibly, grappling with rival hypotheses, and trying to discern which is overall most plausible in light of both direct evidence and background belief. Nor does Wong in his own work by any means relinquish such indefinitively supported claims. Consider Wong's (1996) assertion elsewhere that "powerful socioenvironmental conditions – such as undersocialization, aversive srimulation or deprivation, negative modeling, reinforcement for the sick role, and combinations of these factors…can produce and support psychotic behavior" (p. 321). This etiological claim is made without benefit of the serious evidence appropriate to such a claim, and with the embrace of inferences to unknowns worthy of the worst "unbalanced neurotransmitter" claims Wong criticizes.

Quoting a critic of biological psychiatry, Wong observes:

> Results of over 1,300 outcome studies published since the mid-1950's reveals that 21% of schizophrenic patients on maintenance drug therapy relapse as

> compared to 55% on placebo (Cohen, 1997). Given the bias of scientific journals for publishing positive results (Dickerson, Chan, Chalmers, Sacks, & Smith, 1987) this probably represents an optimistic estimate of the impact of these drugs. By subtracting the relapse rate for neuroleptics from that for placebo treatment, we may compute a "net" drug effectiveness rate of 34%. Undoubtedly, preventing relapse in one third of treated patients is a significant effect; yet, it is probably inadequate justification for past assertions that these drugs were essential or past practices of prescribing these drugs universally. (2006)

Accepting the cited figures, in my view this is a bewildering assessment. We are talking about life-destroying and family-destroying disorders. In any other such medical domain (e.g., cancer treatment), a one-third improvement rate over placebo would be cause for joy, and would be considered justification for recommending universal administration of a therapy, if there is no test that distinguishes the one-third who would benefit (of course allowing for alternative effective therapies and considering patient preferences and other such factors; no treatment should literally be reflexively "universal"). Indeed, much smaller percentages are often reasons for celebration. (As it happens, as I write this there is an article just published that recommends universal chemotherapy, despite its negative side effects, for estrogen-insensitive breast cancers over a certain size, given that a meta-analysis established that 23% of the cases will benefit; the medical literature is replete with smaller benefit percentages that are considered breakthroughs.)

Wong notes that biases in publishing positive results suggest that the effectiveness of medication is less than we know. It is true there are distortions in the system of studying drugs, not only publishing bias but many other issues as well (e.g., selecting proper dosages, large numbers of nonresponders, spontaneous remissions due to the cyclic nature of many disorders, and false positive misdiagnoses) that may lead to complexities of interpretation and inflated apparent effectiveness. However, Wong fails to consider factors that could suggest that drug effects are more powerful than we know. These include, for example, the FDA-mandated "intent to treat" methodology that may not even report the results for those actually completing treatment despite high drop-out rates, the high rate of spontaneous remissions that weaken apparent effectiveness, and the fact that drug trials almost always try one drug at set doses whereas in clinical practice it is well known that individuals often respond very differently to given dosages and to different drugs within a class, and the clinician will often try several different drugs or a combination of several drugs before finding the right prescription that works for a given client (some recent antidepressant studies are starting to take this into account, and seem to be showing higher rates of remission than single-drug studies).

THE PROBLEM OF GENERALIZATION

Wong traces the rise, first, of individualized operant programs, and then token economies for entire wards. The research left no question that patient behavior can be brought significantly under control using behavioral contingencies. True, token effectiveness is by no means universal; some percentage does not respond across studies (Kazdin, 1983). Wong et al. (1987) found that independent recreational activity reduced stereotypic vocalizations, but "that contingent tokens given to one subject for on-task behavior did not contribute significantly to treatment impact" (p. 81), or, as the abstract states, "results were the same with or without contingent tokens" (p. 77), a result mentioned but oddly not pondered in the discussion.

Although Wong's characterization of the success of these approaches is on the optimistic side, the fact is that the data were overall quite positive. So, reasonably, Wong asks: "Why did behavior analysis and related behavioral approaches to treating this disorder fail to develop and thrive? Why did biomedical interventions, particularly psychotropic medication, become the prevailing treatment for severe mental disorders? Is this prevalence based on sound scientific research and technology or something else?"

There is a simple and quite scientific reason why token economies were set aside in favor of drugs; lack of generalization. Drugs could be prescribed for the patient and had roughly the same effect whether inside the institution or out, and outside the institution in principle they can be administered with minimal additional supervision. Behavioral methods depended on continued control over patients of the sort that occurs on a ward, and those environments became rare and some control procedures came to be considered unethical except under exceptional circumstances. In the community, reinforcers cannot be easily controlled, and what control can be achieved is expensive to maintain indefinitely.

That generalization is *the* issue about token economies as applied to most patients has gradually been realized (Glynn et al., 2002; Kazdin, 1982; Kazdin & Bootzin, 1972; Stokes & Baer, 1977). In her much-cited review, Glynn (1990) stated the obvious but often ignored truth: "One of the major challenges for behavior therapy in general (Stokes & Baer, 1977) and token economies in particular (Kazdin, 1982, 1985) is the maintenance and generalization of effects" (p. 387); indeed, "Maintenance and generalization of treatment gains are, of course, critical tests of the utility of treatment interventions" (p. 401). The problem is that "the power of the token economy rests in large part on control of external environmental contingencies," and as the patient moves from the controlled clinic environment to the community and thus can obtain desired rewards through alternative means and avoid punishments thus voluntarily exiting from the token economy, "the treatment modality becomes less useful" (p. 387). Glynn further noted that "the crucial question is, Are token economies primarily prosthetic, resulting in behavioral change only when it is supported by token reinforcement; or are they therapeutic, resulting in long-term behavioral change across settings?" (p. 401), and observed that "The answer to this question has not been fully determined" (p. 401). Nothing has happened since 1990 to change that sober assessment. Glynn does refer to

various factors suggested in the literature that might attempt to address generalization, but these techniques have not been demonstrated to work to maintain gains in open environments after token reinforcement programs for schizophrenics end.

There is a paradox lurking here for behaviorists. As Glynn notes, "the power of the token economy rests in large part on control of external environmental contingencies" (p. 387), and this is demonstrated by the reoccurrence of symptoms when treatment is withdrawn. Yet generalization requires that the target behavior not recur when treatment is withdrawn.

To underscore the reality of the generalizability problem, it is worth citing some of the key empirical literature, without any pretense to a complete or balanced review. In an attempt to replicate Ayllon and colleagues' (Ayllon & Azrin, 1968; Ayllon & Haughton, 1964) classic reports using contingent attention, approval, and token reinforcers to increase and decrease psychotic behavior, Wincze et al. (1972) used token reinforcement to reduce the percentage of delusional verbal in-session behavior of 7 out of 10 treated patients. However, generalization did not occur from the therapy sessions to ward behavior (Liberman et al., 1973 p. 58); Liberman et al. (1972) state that "the effects of both feedback and token reinforcement were quite specific to the environment in which they were applied and showed little generalization to other situations" (p. 247).

Liberman et al. (1973) thus undertook to elaborate on the work with delusional speech by reinforcing rational speech in a multiple baseline design, and using social contingencies to facilitate generalization. They interviewed each of four patients four times per day, ten minutes (potentially) per session; the sessions were ended immediately if the patient talked delusionally, whereas the patients earned time for talking rationally, to be used in an evening chat providing coffee, snacks, and cigarettes. Dramatic changes occurred in rates of delusional talk and rational talk (200-600% increases in the latter) during the sessions (although by no means was delusional talk extinguished; Liberman et al. note that "the treatment reduced delusional speech temporarily and incompletely" [p. 63]), and this generalized somewhat to evening chats (50% reduction in delusional speech in 3 out of 4 patients) where rewards were provided. However, "no generalization of treatment effects could be detected in the routine interchanges that occurred between patients and staff throughout the day outside of the interview sessions and evening chats. The frequencies of delusional remarks remained stable between baseline and Treatment Phase A and increased in Jack during the last treatment phase" (p. 63).

Patterson and Teigen (1973) used an operant conditioning approach to get a psychotic woman who previously had given only delusional answers to direct, factual questions, to answer factually. Operant conditioning trials were continued upon her discharge to a community setting. We know that this sort of intervention can work while the environment remains controlled. During the period in which the treatment continued post-discharge, two interviews were done at 36 and 52 days after discharge outside the sessions, unannounced as evaluations, to assess generalization. Patterson and Teigen say: "No generalization was found in the first interview, but the second gave evidence of some generalization" (p. 65). In fact, in the second interview the patient misled the interviewer (e.g., mentioning her jobs at a hospital without mentioning she was a patient

there), and generally did not respond with clear direct factual answers but rather generally gave "acceptable, though evasive" answers. The authors conclude, "As in previous studies, only limited generalization was obtained" and speculate that "much greater efforts than have previously been employed are necessary to obtain a more desirable level of generalization" (p. 69).

Nelson and Cone (1979) reinforced four categories of behaviors (personal hygiene, personal management, ward work, social skills) in 12 inpatients. There were substantial improvements in the target behaviors, and there was some generalization from the target behaviors to contemporaneous general ward functioning (social competence, neatness, etc.), though no change in manifest psychosis. However, off-ward behavior did not show significant change; indeed, off-ward resident-staff interactions actually went down. Overall, they say that, "Changes…were not dramatic" and that, "as observed changes in off-ward behavior were neither dramatic nor totally consistent in their direction, evidence that treatment gains noted on the ward generalized to off-ward situations is tentative" (p. 268).

Wilder et al. (2001) addressed the same problem and had more success establishing control. But, despite gaining some control over vocalization symptoms, Wilder et al.'s patients reverted roughly to initial levels when treatment was withdrawn. Control was confirmed, but generalization disconfirmed.

MYTH AND REALITY IN THE PAUL AND LENTZ (1977) STUDY

The classic Paul and Lentz (1977) study, cited by Wong, has achieved a deserved iconic status as the most important study of token economy effects on schizophrenia. Moreover, as Dickerson et al. (2005) observe: "There are not any studies since the seminal work of Paul and Lentz (1977) that have formally investigated the transfer of token economy benefits from the hospital setting where the intervention took place to the community." Given that the literature prior to Paul and Lentz tended to raise doubts about generalization of token economy effects, it is important to look carefully at what the Paul and Lentz study did and did not demonstrate.

This was a remarkable study that did indeed support the substantial effectiveness of a token economy as part of a psychosocial training intervention in shaping the behavior of hospitalized psychotic patients. The study evaluated both specific psychotic and general ward behaviors, as well as success of placement in the community as measured by at least 90 days without rehospitalization. The social learning patients ended up changing more, taking less medication, and reaching threshold for being released to the community in greater numbers than those in milieu therapy or traditional hospital care. As Liberman (1980) noted in his review: "The results were astonishing, given the refractory nature of the patients: improved functioning enabling long-term community placement occurred in 97% of the social learning patients with some maintaining themselves for over 5 years which was the longest period of follow-up possible in the study. The milieu therapy program was less effective, but its 71% release and

maintenance rate was still a favorable outcome when compared to the patients treated in the state hospital of whom less than 45% were discharged." (p. 368)

A decade later, Glynn (1990) observed that the social-learning group "spent less time in the hospital, achieved greater discharge rates, were maintained longer in the community, and required less psychotropic medication than either the milieu or traditional-care comparison groups. All differences were statistically significant" (pp. 391-392). Unquestionably, the superiority of the social-learning intervention containing the token economy over milieu therapy or standard hospital treatment in shaping ward behavior was clearly demonstrated.

However, despite intensive intervention over a long time in a controlled environment, the changes were not achieved easily and target behaviors were by no means eliminated. For example, after almost two-and-a half years, "the average social-learning resident had improved to the point of demonstrating normal self-care and interpersonal skills *about half the time*" (p. 425; emphasis added). Nor were all spheres of behavior equally responsive; as Liberman (1980) notes, "Even behaviorists will be disappointed to discover the failure of reinforced sampling-exposure procedures in enhancing these chronic patients' involvement in off-ward, 'therapeutic' activities such as movies, bowling, sewing, games, and a snack bar" (p. 369).

There are two questions I address about the Paul and Lentz study. First, to what degree does the study show that the effects of token economy intervention could generalize to community life? Second, what evidence does the study provide more generally for the stability of behavior changes under environmental variations?

Generalization to the community environment was one of the questions that the study was designed to answer. But we will never know the answer because the follow-up study turned out to be meaningless due to methodological disasters that occurred after release. One problem was that planned follow-up programs and assignment of patients to alternative approaches could not be put in place adequately, due to a variety of circumstances. But this might still have allowed for meaningful follow-up. The first hint of the real problem was that virtually everyone in all treatment groups permanently stayed in the community successfully once they were placed there, so the percentage of success in the community once placed there was about the same virtually perfect rate for all groups (actually, 97%). Glynn observes: "Interestingly, once they had been released, patients from all three conditions had comparable rates of successful placement in the community. This result suggests that, once discharge readiness and aftercare placement have been achieved, other variables independent of the predischarge treatment have more influence on community tenure" (p. 392). That is, because virtually everyone who was released from any of the groups stayed released, the greater number of social-learning patients who reached behavioral threshold for release implied a greater number who stayed released, but the reasons for their staying released are likely different from the reasons that got them to release threshold in the first place. As we shall see, the non-treatment factor to which Glynn alludes can likely be precisely identified as medication.

Second, a sizable percentage of released patients, and about the same proportion of each group, deterioriated in their behavior from each of the three treatment groups,

constituting about 32% in each of the groups (Paul & Lentz, p. 407). Interestingly, many of these, 21% altogether, fell below where they had been when earlier rejected for aftercare and thus were below the threshold to get placed in the community, yet were not returned to the ward. Clearly, as Paul and Lentz observe, the community facilities' criteria for staying in release were more liberal than their criteria for bringing a new patient into the facility.

It thus became apparent that once an individual was in the community, the standards were stretched to keep them in the community. This raises the question of why the benchmark for release was not lower to begin with. If it had been lower, more individuals from the various treatment groups might have been released successfully, lowering the relative success of the social learning group. Paul and Lentz do not analyze how many individuals might have been released in retrospect if these lower levels of reshaped behavior were sufficient to warrant release to the community.

The third and by far the most devastating problem for drawing any scientific conclusions about generalization from the Paul and Lentz study—and the most revealing for understanding the surprising community results—concerned indiscriminate post-release medication. During the intramural period many patients—and especially the social learning patients—had had their medication eliminated or reduced (an important finding in itself regarding intramural success). However, when patients from any of the treatment groups were released to the community, irrespective of their earlier medication status, they were seemingly routinely placed on medication. In one of the two main community settings, about 80% of the patients were found to be on medication at all three six-month-apart follow-up interviews, and at the other setting, 80% were found to be on medication during at least two out of three of the follow-up interviews (Paul & Lentz, p. 397). The massive use of psychotropic medication for released patients essentially wiped out the chance to evaluate the generalizability of the behavioral program in itself. And, it likely explains the universal success of all patients in remaining in the community; the overwhelming success of all groups was based on the medication. Between retaining regressed patients and medicating almost everyone, the community component as a test of generalizability was thoroughly compromised. As Paul and Lentz themselves note in considering generalizability, "[T]he indiscriminant use of psychotropic drugs, severely limits any other conclusions" (p. 410). They suggest that the use of medication may have interfered with what would have been a post-release transfer of learning from the token economy group (p. 441), and recommend less automatic use of drugs in release facilities; but whether the drugs did indeed interfere with what would have been a documented superior transfer of learning in the social-learning approach cannot be known.

Granting that the Paul and Lentz study fails to demonstrate (or even to test) generalizability, what does it say about the stability of token-economy gains over the 4.5 years of the intramural part of the study, surely one of the most sustained tests of a token economy for mental disorder? The answer appears to be that due to an unexpected natural experiment, the study inadvertently provides powerful evidence of the instability of behavioral gains. (In what follows, it should be kept in mind that although for simplicity I write as if there is an identifiable constant population being treated over time within each

group, this is only approximately true, because at intervals some patients were released to the community as they reached behavioral thresholds and new patients introduced to the program.)

Essentially, what happened is this: Aggressive behavior on the ward was initially very heavily punished with 72 hours of time-out in a solitary room (Liberman [1980] notes that this is "a duration that is incompatible with current guidelines on human rights" [p. 369]) plus heavy token penalties. The staff soon perceived that some aggressive patients seemed to be seeking punishment because of the reward of the quiet of the timeout room allowing for sleep rather than participation in the ongoing activities of the ward. They thus initiated actions to make the timeout room a more aversive experience, for example, making it hot and humid and blaring sound at the patient periodically to interrupt naps. Other techniques to control aggressiveness included "High-dose neuroleptic drugs used as "chemical straight-jackets"; two-way telereceivers; part-time male college students hired to study at night on the units; and even beefed-up security controls" (Liberman, 1980, p. 369), and even tear gas was purchased and considered and then abandoned. The most satisfactory outcome occurred when male staff were assigned for a greater number of hours so female staff were protected.

It turned out that the success of the entire ongoing program was highly sensitive to the availability of severe forms of aversive stimulation as penalties. The initial 72-hour period was lowered to 2 hours due to changing administrative rules, later increased to 24 hours, and eventually returned to 72 hours. Only at the 72-hour level did the token economy treatment display continued outcome improvement and effectiveness.

Liberman (1980) observed:

> Even in the token economy, only a minimum of 72 hours of time out seemed to control aggression – a duration that is incompatible with current guidelines on human rights….Paul and his team reluctantly encountered a natural experiment with a withdrawal design, finding a tremendous increase in "intolerable behavior" when the duration of permissible time out was reduced by administrative fiat to 2 hours. Even when this limitation was rescinded and up to 24 hours of time out was allowed, the average weekly incidence of aggressive acts remained above that occurring during the baseline period before the token economy was begun!...
>
> During the period when time out was limited to 2 hours and aggression markedly increased, the continuous data collected on the units revealed a serious regression among the patients on both psychosocial programs in all levels of performance. In fact, during this period patients in the milieu program experienced a washout of almost all the gains they had acquired since the start of the project. During the last 6 months of the project when time out was again lengthened, patients on both programs again showed progressive improvements in self-care, interpersonal skills, instrumental role performance, and bizarre behavior. (pp. 369-370)

Thus, "during the next year and a half (anniversary assessments 5 to 8) when shortened lengths of expulsion time and time out for assaultive behavior were in effect in the psychosocial programs, both showed significant decreases in IAB functioning" (p. 375). The problem here was *not* regression after treatment was withdrawn; it was regression as treatment continued as before for a wide variety of behaviors, with the only change being the rules for dealing with aggressive behavior. Consider the magnitude of this regression: At the seventh and eighth six-month assessments taking place just twelve and six months before the end of the intramural part of the study (this is at 3.5 and 4.0 years after intervention initiation!), there was, according to Paul and Lentz: (1) no significant change in the Inpatient Assessment Battery (IAB) scores of the social learning group from the level of functioning at study initiation ("The social-learning group failed to show a significant increase from the level before the program introduction only at anniversary assessments 1, 7, and 8" [pp. 374-375]); (2) no significant difference in the overall IAB assessment measure between the social-learning and traditional-hospital-care components ("Mean IAB Functioning of the social-learning group was significantly…greater than that of the hospital group through assessment 6" [p. 373])); and (3) no significant difference between the social-learning and hospitalization groups in the changes in IAB levels from program initiation to these points in time ("The mean increase of the social-learning group were also significantly better than the changes of the hospital group at every anniversary assessment except 7 and 8, where the groups did not differ significantly" [p. 373]). The changes in mean IAB scores of the three study groups, measured every six months over the four-and-a-half years of the intramural part of the study, are graphically represented in Paul and Lentz's crucial Figure 34.1 (p. 374), and that Figure reveals dramatic decreases in functioning in the social-learning group after the time-out period was lowered, and virtual convergence statistically at three-and-a-half and four years of the social-learning and routine hospital care groups.

[NOTE TO THE READER: I apologize for the fact that the relevant Figure 34.1 from Paul and Lentz, which I expected to accompany and clarify the text, could not be presented here because Dr. Paul refused me permission to reprint this figure from his 1977 book. I encourage the reader to consult the original book. One might argue that failure to grant such permissions impedes open debate and the fair airing of scientific arguments, and thus illustrates the kind of intrusion of unscientific use of power into scientific discourse that this interchange is considering.]

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. The basic point stands that, even after years of intensive intervention, overall progress on basic behaviors and on clinically appropriate and inappropriate behaviors was highly sensitive to a single change such as lowered punishment for aggressive behavior. A spurt in the last six months of a five-year program

due to reinstitution of Draconian punishment for aggression basically reshaped the results in the nick of time: "During the last six months of the intramural period, following the reintroduction of procedures and the changes for controlling assaultive behavior, both psychosocial programs showed sharp increases in IAB Functioning (see Figure 34.1)" (p. 375). As a glance at Paul and Lentz's Figure 34.1 (the one for which permission to reprint was unfortunately not granted) would have revealed, there was no remotely proportional dose-response relationship over time; at 4 years, the outcome results would have been utterly different in terms of effectiveness in changing behavior than they were at 4.5 years.

We can now return to a hypothesis implicitly suggested by Wong. We saw that Wong suggested that the problem with extinguishing overlearned psychotic symptoms is that patterns reinforced over a lifetime cannot be dealt with in the weeks or months of typical behavioral intervention. Translated into a prediction, Wong is saying that the patterns could be extinguished with sufficient time to allow overlearning to be undone. The Paul and Lentz study appears to offer a disconfirmation of Wong's implied claim. Even with lengthy treatment, behavioral control could be maintained only with continued high levels of punishment threat.

In the present article, Wong traces the triumphs of token economies through several stages, then laments the waning of this approach. But elsewhere, oddly enough, Wong (1996) himself, enamored of the idea of individualized functional analysis as an alternative to generic token economies, makes the very same point I am making about the limitations of such approaches:

> Although studies have suggested that the aforementioned interventions have had favorable outcomes, this technology is becoming obsolete. Investigators…imposed arbitrary reinforcers and reductive consequences to override inappropriate responses; however, they did so without first conducting a functional analysis that would have identified the preexisting reinforcement contingencies causing the problem behavior….And even if arbitrarily imposed consequences are potent, therapeutic gains may rapidly disappear as soon as the contrived reinforcers and consequences are withdrawn. (pp. 327-328)

In other words, token economies are "obsolete" because they failed to generalize post-treatment; they failed to generalize because they did not directly address the reinforcers that are the cause of psychosis (for this general causal claim there was and is of course no evidence whatever), and this may be remedied by individualized functional analysis. This was written in 1996. Now, ten years later, no one has demonstrated that functional analysis can enduringly change schizophrenia where token economies failed. And the reality Wong observed about the lack of generalization of token economy outcomes surely has not changed. Wong's change of emphasis does not appear to be based on a balanced scientific assessment of the evidence.

CONCLUSION

Behaviorist research on treatment of schizophrenia can boast legitimate achievements that are worth defending. Clearly, irrespective of the generalization issue, when ward behavior itself is what is desired to change, token control is very useful indeed. Thus, when dealing with treatment-refractory patients who cannot sustain enough progress to be released or who are problematic on the ward, this can be an appropriate treatment, if modern ethical guidelines can be met.

However, in their zeal to revive a flagging expansive behavioral ideology that goes beyond what is scientifically plausible, Wong, Wyatt, and Midkiff seem to be deploying all the techniques of which pseudoscientists have been justly accused in the past. These include not taking into account the entire gamut of evidence bearing on an issue, insisting on the existence of as-yet-unconfirmed causal processes that fit one's theory but have no prima facie plausibility, citing politics and other unscientific considerations in defending one's position, invoking various methodological factors such as parsimony rather than addressing the evidence, abandoning parity of reasoning in assessing rival theories, using verbal gymnastics to preserve the semblance of cherished principles semantically even where the substance has been disconfirmed, simply not allowing disconfirmation or accepting reasonable plausibility arguments, and so on.

Others recognize and are attempting to constructively address the real limitations of behavioral evidence. For example, Glynn et al. (2002) state at the outset: "Although skills training is a validated psychosocial treatment for schizophrenia, generalization of the skills to everyday life has not been optimal" (p. 829). Their modestly successful treatment, however, continues to intervene with the patient in an intensive way in the community until the end of the study. Thus far, the possibilities and limits of generalization after treatment termination or with minimal maintenance have yet to be established. This scientific failure cannot be laid at the feet of biological psychiatrists. Wong, Wyatt, and Midkiff's strategy of trying to save behaviorist approaches to schizophrenia by invoking an ad hoc auxiliary hypothesis, that the eclipse of token economies is all a matter of power and not a matter of science, sadly takes behaviorists one step further away from confronting the real scientific issues that beset them.

REFERENCES

Ayllon, T., & Haughton, E. (1964). Modification of symptomatic verbal behavior of mental patients. *Behavior Research and Therapy*, *2*, 87-97.

Ayllon, T., & Azrin, N. H. (1968). *The token economy: A motivational system for therapy and rehabilitation*. Englewood Cliffs, NJ: Prentice-Hall.

Baker, R., Hall, J. N., Hutchinson, K., Bridge, G. (1977). Symptom changes in chronic schizophrenic patients on a token economy: A controlled experiment. *British Journal of Psychiatry, 131,* 381-393.

Dickerson, F. B., Tenhula, W. N., & Green-Paden, L. D. (2005). *Schizophrenia Research*, *75*, 405-416.

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

Glynn, S. M., Marder, S. R., Liberman, R. P., Blair, K, Wirshing, C. W., Wirshing, D. A., Ross, D., & Mintz, J. (2002). Supplementing clinic-based skills training with manual-based community support sessions: Effects on social adjustment of patients with schizophrenia. *American Journal of Psychiatry*, *159*, 829-837.

Kazdin, A. E. (1982). The token economy: A decade later. *Journal of Applied Behavior Analysis, 15,* 431-445.

Kazdin, A. E. (1983). Failure of persons to respond to the token economy. In E. B. Foa & P. Emmelkamp (Eds.), *Failures in behavior therapy* (pp. 335-354). New York: Wiley.

Kazdin, A. E., & Bootzin, R. R. (1972). The token economy: An evaluative review. *Journal of Applied Behavior Analysis*, *5*, 343-372.

Keller, M. C., & Miller, G. (in press). Resolving the paradox of common, harmful, heritable mental disorders: Which evolutionary genetic models work best? To appear in *Behavioral and Brain Sciences*.

Liberman, R. P., Teigen, J., Patterson, R., & Baker, V. (1973). Reducing delusional speech in chronic paranoid schizophrenics. *Journal of Applied Behavior Analysis*, *6*, 57-64.

Liberman, R. P. (1980). A review of Paul and Lentz's Psychological Treatment for *Journal of Applied Behavior Analysis*, *13*, 367-371.

Nelson, GL, and Cone, JD Multiple-baseline analysis of as token economy for psychiatric inpatients. (1979). *Journal of Applied Behavior Analysis*, *12*, 255-271.

Patterson, RL, and Teigen, JR (1973) Conditioning and post-hospital generalization of nondelusional responses in a chronic psychotic patient. *Journal of Applied Behavior Analysis*, *6*, 65-70.

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

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis, 10,* 349-367.

Wakefield, J. C. (1992). The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist*, *47*, 373-388.

Wakefield, J. C. (1998). The DSM's theory-neutral nosology is scientifically progressive: Response to Follette and Houts. *Journal of Consulting and Clinical Psychology*, *66*, 846-852.

Wakefield, J. C. (1999a). Evolutionary versus prototype analyses of the concept of disorder. *Journal of Abnormal Psychology*, *108*, 374-399.

Wakefield, J.C. (1999b). The concept of mental disorder as a foundation for the DSM's theory-neutral nosology: Response to Follette and Houts, Part 2. *Behavior Research and Therapy*, *37*, 1001-1027.

Wincze, J.P., Leitenberg, H., and Agras, H. S. (1972). The effects of token reinforcement and feedback on the delusional verbal behavior of chronic paranoid schizophrenics. *Journal of Applied Behavior Analysis*, *5*, 247-262.

Wilder, D. A., Masuda, A., O'Connor, C., & Baham, M. (2001). Brief functional analysis and treatment of bizarre vocalizations in an adult with schizophrenia. *Journal of Applied Behavior Analysis*, *34*, 65-68.

Winkler, R. C. (1970). Management of chronic psychiatric patients by a token reinforcement system. *Journal of Applied Behavior Analysis*, *3*, 47-55.

Wong, S. E., Terranova, M. D., Bowen, L., Zarate, Roberto, Massel, H. K., & Liberman, R. P. (1987). Providing independent recreational activities to reduce stereotypic vocalizations in chronic schizophrenics. *Journal of Applied Behavior Analysis*, *20*, 77-81.

Wong, S. E. (1996). Psychosis. In M. A. Mattaini & B. A. Thyer (Eds.), *Finding solutions to social problems: Behavioral strategies for change* (pp. 319-343). Washington, DC: American Psychological Association.