Adm Policy Ment Health (2008) 35:283–294
DOI 10.1007/s10488-008-0174-y

ORIGINAL PAPER

When Programs Benefit Some People More than Others: Tests of Differential Service Effectiveness

Cathaleene Macias · Danson R. Jones · William A. Hargreaves · Qi Wang · Charles F. Rodican · Paul J. Barreira · Paul B. Gold

Published online: 30 May 2008

Abstract Practitioners need to know for whom evidence-based services are most or least effective, but few services research studies provide this information. Using data from a randomized controlled comparison of supported employment findings for two multi-service psychiatric rehabilitation programs, we illustrate and compare procedures for measuring program-by-client characteristic interactions depicting differential program effectiveness, and then illustrate how a significant program-by-client interaction can explain overall program differences in service effectiveness. Interaction analyses based on cluster analysis-identified sample subgroups appear to provide statistically powerful and meaningful hypothesis tests that can aid in the interpretation of main effect findings and help to refine program theory.

C. Macias (✉) · D. R. Jones · Q. Wang · C. F. Rodican
McLean Hospital, Belmont, MA, USA
e-mail: cmacias@mclean.harvard.edu

D. R. Jones
e-mail: dansonjones@sbcglobal.net

Q. Wang
e-mail: qi.wang@eeoc.gov

C. F. Rodican
e-mail: crodican@mclean.harvard.edu

W. A. Hargreaves
University of California, San Francisco, CA, USA
e-mail: billharg@comcast.net

P. J. Barreira
Harvard University, Cambridge, MA, USA
e-mail: pbarreira@uhs.harvard.edu

P. B. Gold
Medical University of South Carolina, Charleston, SC, USA
e-mail: pbg2006@verizon.net

Keywords Treatment moderation · Co-occurring disorders · Subgroup analyses · Supported employment

Introduction

Services research findings are most useful to practitioners when they specify the type of person for whom an intervention has been found to be effective (Chambless and Hollon 1998; Kraemer et al. 2002; Miklowitz and Clarkin 1999; Rothwell 2005; Wells 1999). Findings become specific when a research sample is precisely defined or limited to a single diagnostic group, as is the case in many therapy and pharmaceutical trials (e.g., Jensen et al. 1999; McBride et al. 2006). However, heterogeneous samples are also needed to test the efficacy of treatments for complex patients who do not fit clearly into standard demographic or diagnostic categories (Ackerman 1999; Blankertz 1998; Ruscio and Holohan 2006). Sample homogeneity precludes heterogeneity, and yet common sense suggests that a blend of these two approaches would be advantageous for assessing the effectiveness of an intervention for different types of people.

Homogeneous samples, by definition, have minimal variability, so homogeneity increases confidence in findings (internal validity) by minimizing the likelihood that intervention outcome differences can be explained by differences in participant characteristics. However, most services research studies need large heterogeneous samples to maximize generalizability (external validity), so services researchers are legitimately concerned that variation in sample characteristics across experimental conditions could pose a threat to the validity of findings. Random assignment will not assure equitable allocations of characteristics across experimental groups unless a sample is very large (Krause

and Howard 2003). To safeguard internal validity, services researchers typically compare intervention samples on commonly measured characteristics, and then statistically control for any detected differences. If a control variable is a significant predictor of outcomes, this means that participants with this particular characteristic had overall better (or worse) outcomes, assuming that all other measured characteristics are held constant. Statistically controlling for outcome-related sample characteristics does not identify for whom each intervention was most effective, but it does allow researchers to assume that sample differences between experimental conditions did not account for overall intervention differences in outcomes.

Unfortunately, the statistical control of participant characteristics can instill false confidence in the validity of research findings, curtailing a search for other possible alternative explanations. Differential effectiveness can also compromise internal validity if it is not taken into account. This is especially true of null findings because an absence of overall significant differences in intervention outcomes can mask the fact that one service was more effective for certain clients, while the comparison service was more effective for others (e.g., Bickman et al. 1999; King et al. 2000). Likewise, it is important to check whether differential effectiveness could explain statistically significant differences in outcomes whenever one experimental intervention is designed to benefit some clients more than others, and that type of client is prevalent (or underrepresented) in the total study sample (Bühringer 2006). For instance, if one experimental program has medical staffing, while a comparison program does not, we would expect the program with medical staff to have better health-related outcomes if the study sample has many individuals with physical health problems. Even if random assignment creates comparable experimental groups, a high percentage of unhealthy individuals in both conditions would favor the medical program. Likewise, screening out applicants with health problems during recruitment would give the non-medical program an unfair advantage if health problems were prevalent in the study population. Neither comparability in sample characteristics across experimental conditions, nor the statistical control of detected sample differences, is sufficient to ensure internal validity. Main outcome findings can also be explained by variations in service effectiveness that reflect *with whom* and *how* each program was intended to be effective.

Methods for Testing for Differential Service Effectiveness

A preferred method for testing whether a service was more or less effective for particular types of clients is to enter variable-by-intervention interaction terms as predictors in a multivariate analysis of variance or 'moderated multiple regression' analysis (Aiken and West 1991; Stone-Romero and Anderson 1994). For instance, one program might be expected to be more effective than another program for older, physically unhealthy people, and this hypothesis could be tested by adding 'age-by-intervention' and 'health-by-intervention' interaction terms to the analysis. To test the more specific hypothesis that a program was less effective for older individuals with health problems, an 'age-by-health-by-intervention' interaction term would also be needed.

Alternatively, a heterogeneous sample could be disaggregated into relatively homogeneous subgroups based on participant characteristics expected to moderate intervention effectiveness (e.g., King et al. 2000; Pettinati et al. 2000; Uehara et al. 2003). For instance, a sample might be disaggregated into subgroups defined by age and health, so that individuals with commonly associated characteristics are grouped together within each experimental intervention (e.g., 'older, unhealthy subgroup-by-intervention' interaction term). Subgroups can be defined using variable cut-points (e.g., median scores), ranked categories, or category combinations (e.g., older women). Alternatively, if a sample is large, statistical techniques, such as cluster analysis, can be used to group individuals who share the same constellation of characteristics.

Variable and subgroup-based analyses are both viable strategies for testing hypotheses about differential service effectiveness when a sample is relatively homogeneous, and/or individuals fall into distinct, comparably sized groups based on one or two key variables related to intervention success. When a sample is very heterogeneous, and especially when each individual can be characterized by several related characteristics, subgroup analyses would appear to be more meaningful and statistically powerful than analyses based on variables (Aguinis and Stone-Romero 1997). This is because complex individuals who have co-occurring characteristics must be depicted using complex 'higher order' interaction terms (e.g., 'age-by-health-by-substance use'), each of which requires the additional inclusion of not only the main variables (e.g., age, health, substance use), but also 'lower order' interaction terms that together represent all possible variable combinations (e.g., 'age-by-health,' 'age-by-substance use,' 'health-by-substance use'). By contrast, a single interaction term is sufficient for depicting this same level of complexity in a subgroup-based analysis ('older age, poor health, low substance use' versus all participants without this combination of attributes), and any number of unique subgroups can be compared as long as each individual is assigned to a single subgroup. For this reason, subgroup analyses appear to be particularly advantageous for service programs designed to serve individuals who have multiple co-occurring disorders or dual diagnoses (Batstra

et al. 2002; Bekker 2003; Beutler et al. 1996; Kraemer et al. 2001).

Role of Program Theory in Hypothesis-formulation

Tests of differential effectiveness should always be program-specific and designed to refine program theory, rather than pursued through exploratory analyses. Atheoretical analyses that rely on trial-and-error explorations, and statistical methods that capitalize on covariation (e.g., stepwise regression), will almost always identify one or more types of client who did especially well or poorly in a particular intervention, but these findings will very likely be due to chance alone. Hypotheses derived from program theory will yield more practical and valid insights into service effectiveness because they specify and limit the number of planned analyses. Fortunately, service manuals and intervention descriptions abound with assumptions about who should benefit most and why, and these assumptions are easily translated into testable hypotheses prior to data analysis (Howell and Peterson 2004; Stout and Hayes 2005; West and Aiken 1997).

Overview of Present Study

In this article, we use an existing dataset collected for a randomized controlled trial of supported employment to compare the relative sensitivity of four methods of testing for differential service effectiveness: (1) continuous variables, (2) categorical variables, (3) subgroups based on categorical variables, versus (4) cluster analysis-identified subgroups. We then reinterpret our study's previously published main findings (Macias et al. 2006) in light of these post hoc subgroup analyses to illustrate how tests of predicted variations in service effectiveness can help to refine program theory. In our example, we pay close attention to the relative effectiveness of our two experimental programs for providing supported employment services to adults with severe mental illness who also have physical health problems and/or substance use disorders that might limit job attainment. One intervention was a vocationally integrated program of assertive community treatment (PACT; Allness and Knoedler 1998; Frey 1994; Stein and Test 1980), which is a mobile team providing out-of-office psychiatric care, help with daily living, crisis intervention, substance use treatment, and medical care, in addition to supported employment services. The other intervention was a facility-based clubhouse modeled on Fountain House in Manhattan (Anderson 1998; Beard et al. 1982) that, in keeping with international clubhouse standards (Propst 1992), provided no medical services or substance use treatment, but offered case management, social support, supported housing, supported education, supported employment, and a workplace environment designed to encourage members to relinquish a patient identity and return to a normal life (Propst 1992). Based on these service model characteristics, we hypothesized in our original application for grant funding that PACT would be most vocationally effective for adults with severe mental illness who had chronic physical health problems or severe substance use, whereas the clubhouse would be most effective for those who were relatively healthy with no severe substance use.

Methods

Study Design

Data for these analyses were from a long-term services evaluation project conducted in Worcester, Massachusetts from 1996 to 2001 (Macias et al. 2006). This randomized controlled trial assigned adults with serious mental illness ($N = 177$) to one of two community-based psychiatric rehabilitation interventions following procedures approved by the McLean Hospital IRB. In both multi-service programs, staff trained in supported employment (Bond et al. 2001; Trach 1990) worked closely with other staff to ensure rapid placement of participants into mainstream jobs not reserved by employers for individuals with disabilities.

Sample Description

Study applicants were recruited in 1996–1998 from 42 local organizations, and through posted flyers, radio, and newspapers. Any individual over age 18 was eligible if she or he were unemployed and had a clinician diagnosis of schizophrenia spectrum disorder, bipolar disorder, or recurrent severe depression, but no diagnosis of severe mental retardation. One enrollee crossed-over to the unassigned service, and employment data were unavailable for two others. The remaining sample ($N = 174$) was heterogeneous and similar to larger epidemiological samples within the same state (Jones et al. 2004) in demographics and health problems (Dickey et al. 2002), as well as in mortality rate (Dembling et al. 1999).

Work-related Grouping Variables

Investigation of PACT and clubhouse differences in vocational effectiveness focused on four potentially disabling factors known to limit employment in psychiatric populations: *psychiatric symptom severity* (Anthony et al. 1995; Chwastiak et al. 2006; Goldberg et al. 2001; Razzano et al. 2005; Regenold et al. 1999; Slade and

Salkever 2001), *physical health problem severity* (Dixon et al. 2001; Druss et al. 2000; Razzano and Hamilton 2005; Razzano et al. 2005), *older age* (Burke-Miller et al. 2006; Cook et al. 2001; Goldberg et al. 2001; Mueser et al. 2001; Wewiorski and Fabian 2004), and *substance use* (Lehman et al. 2002; Razzano et al. 2005). Gender is not predictive of work among adults with serious mental illness (Burke-Miller et al. 2006); ethnicity was restricted in this Massachusetts study sample (98% Caucasian).

These four client variables were measured concurrently with employment across the 1996–2001 data collection period (rather than only at baseline) to allow identification of chronic conditions that might continuously or sporadically prevent or hinder employment, including the onset or worsening of conditions after study enrollment (Batstra et al. 2002; Kraemer et al. 2006). Health conditions, psychiatric symptoms, and substance use tended to be persistent, with no discernable patterns of temporal variation after first incidence that would suggest service-related changes may have mediated program outcomes.

Psychiatric symptoms were measured as total scores on the Positive and Negative Syndrome Scale (PANSS; Kay et al. 1987) averaged across all interviews completed during each participant's first 30 months in the project (median: 6 scores); subscale scores were equally weighted. Interviewers were trained by Lewis Opler, MD and had high inter-rater reliability (Salyers et al. 2001). Physical health problems were identified through open-ended PANSS probes, as well as Medicaid claims and interviewer observations, and each chronic or permanent health problem was assigned the least severe ICD-9 diagnostic code that fit the medical description. Each condition was then coded for severity using the Chronic Illness and Disability Payment System, which is based on actual treatment costs for a large multi-state sample of Medicaid recipients (Kronick et al. 2000). Physical health problem severity scores were the sum of estimated annual costs for each participant's most costly physical condition within each of 14 diagnostic categories (Jones et al. 2004). Substance use disorders were identified through clinician reports, interviews, and treatment records, and coded 0 (minimal or none), 1 (moderate), or 2 (severe). A moderate rating indicates any clinician report of severe dependence or treatment lasting more than 5 days; a severe rating indicates recurrent, life-disrupting substance abuse.

Methods for Disaggregating the Total Sample into Independent Subgroups

In addition to studying variations in PACT and clubhouse work outcomes for participants high and low on each of these four client characteristic variables, we divided the sample into subgroups that took into account co-variation in the four characteristics. Two methods were compared.

Median Splits on Grouping Variables We first created sample subgroups by dichotomizing each of the four grouping variables based on median scores, and examining cross-tabulations for these variable groupings. The intent was to assign each individual to a specific category, so that subgroups would be independent. Our procedures were admittedly arbitrary, but logical, and we set a goal of at least 30 individuals per group. We first examined the four categories created by a cross-tabulation of age (older, younger) and physical wellness (healthy, unhealthy) categories. Only 4 individuals fell into the older age, healthy category, so we placed these 4 into the younger, healthy category and relabeled it *healthy* ($n = 35$). The remaining *older, unhealthy* ($n = 38$) category was adequate in size, but the younger, unhealthy category was large enough ($n = 101$) to disaggregate based on the other two grouping variables, substance use and psychiatric symptoms. There was a low positive association between youth and substance use, so we created *high substance use* ($n = 41$) versus low or no substance use ($n = 60$) subgroups within the younger, unhealthy category. We then further divided the young, unhealthy, low substance use subgroup based on the median score for psychiatric symptoms, creating a *young, psychiatrically ill* subgroup ($n = 30$) that was low on substance use and a *young, physically ill* subgroup ($n = 30$) that was low on both substance use and psychiatric symptoms. ANOVA validation tests confirmed that these five subgroups differed ($P < .001$) in ways reflected in the subgroup labels.

Cluster-analysis Following the examples of James et al. (2006) and Peck (2005), we also identified subgroups using cluster analysis because this statistical strategy would generate homogeneous groupings based on the natural co-occurrence of the four characteristics (Batstra et al. 2002; Rapkin and Dumont 2000). We used a Ward procedure (1963) and the hierarchical agglomeration technique with squared Euclidean distances (SPSS 1999). The cluster analysis identified five subgroups: *very psychiatrically ill* ($n = 35$), *very physically ill* ($n = 31$), *substance use disorder* ($n = 31$), *older, chronically physically ill* ($n = 25$), and *relatively healthy* ($n = 52$). As with the subgroups based on variable median splits, ANOVA validation tests confirmed that these five subgroups differed on each of the four client variables at $P < .001$.

Employment Rates

Employment was operationally defined as any job lasting at least 5 days that met the US Department of Labor's definition of competitive employment: mainstream, integrated

work paying at least minimum wage (Department of Labor 1998; Workforce Investment Act of 1998). Clubhouse transitional employment met these criteria, but we did not count TE as an outcome so that our findings would be comparable to the findings of other supported employment studies. The two programs kept identical employment records, which were corroborated by self-report data collected during 6-month and final exit interviews, as well as telephone calls to family members.

Control Variables

Program Preference

To control for participants' pre-existing attitudes toward either experimental program (Macias et al. 2005), we recorded each applicant's program preference at the time of application, and then recoded these preferences as match and mismatch to preference versus no prior preference following randomization to experimental conditions.

Work Interest

Participants' stated interest in work (1, yes; 0, no or uncertain) was measured during the first research interview, prior to randomization. Work interest is a strong predictor of employment (Drebing et al. 2005; Macias et al. 2001; Regenold et al. 1999) and often used as a screening criterion by supported employment programs.

Receipt of Employment Services Total hours of help with job searches (logged) were derived from daily service logs kept by all staff from January 1996 through December 2000.

Data Analysis Plan

We tested our research hypotheses using moderated multiple regression (Aguinis 2004), a preferred method for subgroup comparisons (Aiken and West 1991; Lipchik et al. 2005) and risk-adjustment (Hendryx et al. 2001; Hendryx and Teague 2001). Program-by-client characteristic interaction terms (Judd and Kenny 1981; Kenny et al. 2004) were created by multiplying each participant's variable category (1, high; 0, low), or centered variable score (Aiken and West 1991), by program assignment (1, PACT; 0, clubhouse). Subgroups were compared as dichotomous variables (1, membership in the subgroup; 0, membership in another subgroup), with one of the five subgroups serving as the reference category in each analysis. To control for multiple tests, we conducted hierarchical regression analyses (SPSS 1999) and required each block of conceptually similar variables to reach statistical significance as an omnibus test before interpreting any significant beta within the block.

Results

Preliminary Analyses

T-tests revealed experimental program differences on two of the four key variables expected to moderate service effectiveness: PACT clients had worse (higher) psychiatric symptoms, while clubhouse clients had more severe physical health conditions. Because both variables have correlated negatively with employment in previous studies, the difference in psychiatric symptoms favored the clubhouse, while the difference in physical health favored PACT. The only significant correlation between the four variables was for age and physical health ($r = +.29$, $P < .01$). Older individuals tended to be in poorer health. We rephrased our hypotheses to take this correlation into account: PACT should be most vocationally effective for older clients with health problems, and for clients with severe substance use disorders. The clubhouse should be most effective for clients who are younger and relatively healthy without severe substance use.

Aim I. Comparison of Four Methods for Calculating Client-by-program Interaction Terms

We conducted four separate regression analyses (Table 1), each of which measured the four key client characteristics in a specific way. Analysis 1 used continuous variable scores on the four client variables, with health condition severity log-transformed. In Analysis 2, these four variables were dichotomized based on median splits, with the age-by-health interaction term representing a four-category variable. Since every individual was grouped as either high or low on each variable, these groupings were not independent. Analysis 3 compared independent subgroups that were identified through cross-tabulations of the four dichotomized variables used in Analysis 2. In Analysis 4, we identified five independent and homogenous subgroups using cluster analysis.

Statistical Sensitivity

We compared the relative sensitivity of each method for testing our hypothesis that older adults with chronic health problems would have a higher employment rate if assigned to PACT. The older, unhealthy subgroup was the reference group in Analyses 3 and 4.

Table 1 presents statistics for the predictor variables in each block at the time the block was entered into the

Table 1 Comparative sensitivity of four measurement methods for detecting a program-by-subgroup interaction effect ($N = 174$)

	β	SE	P		β	SE	P
Analysis 1: continuous variables[a]				*Analysis 3: median-split subgroups; reference = older, unhealthy*[a]			
Block 1: program (PACT)	.79	.31	.011	Block 1: program (PACT)	.79	.31	.011
Block 2: client characteristics				Block 2: client characteristics			
Psychiatric symptoms	−.39	.17	.023	Psychiatrically ill	.13	.51	.790
Substance use	.21	.16	.188	Severe substance use	.26	.47	.575
Age	−.22	.17	.197	Psychiatrically well	.27	.50	.586
Physical health	−.06	.18	.718	Physically well	.22	.49	.648
Age × physical health	.09	.16	.584	Block 3: program-by-characteristics			
Block 3: program-by-characteristics				Program × psychiatrically ill	–1.21	1.06	.252
Program × psychiatric symptoms	.04	.34	.905	Program × substance use	–1.26	.98	.200
Program × substance use	−.42	.33	.205	Program × psychiatrically well	–2.56	1.06	.015
Program × age × physical health	.84	.50	.094	Program × physically well	–1.10	1.03	.286
Analysis 2: dichotomized variables[a]				*Analysis 4: cluster-analysis subgroups; reference = older, unhealthy*[a]			
Block 1: program (PACT)	.79	.31	.011	Block 1: program (PACT)	.79	.31	.011
Block 2: client characteristics				Block 2: client characteristics			
High/low psychiatric symptoms	−.64	.33	.052	Psychiatrically ill	−.98	.56	.081
High/low substance use	.40	.33	.228	Severe substance use	−.08	.55	.890
Older/younger age	–1.42	1.23	.250	Very physically ill	−.22	.56	.694
Better/worse health	−.27	.43	.527	Relatively healthy	.41	.50	.414
Older/younger × better/worse health	1.18	1.29	.362	Block 3: program-by-characteristics[b]			
Block 3: program-by-characteristics				Program × psychiatrically ill	–3.31	1.33	.013
Program × psychiatric symptoms	.06	.67	.925	Program × substance use	–1.42	1.37	.301
Program × substance use	1.26	.70	.073	Program × very physically ill	–3.64	1.37	.008
Program × age × physical health	1.86	.86	.031	Program × relatively healthy	–2.81	1.23	.022

[a] Full regression models: analysis 1: $\chi^2 = 19.70$, df = −9, $P < .020$; −2 log likelihood = 220.4, Nagelkerke $R^2 = .143$; Analysis 2: $\chi^2 = 21.01$, df = 9, $P < .013$; −2 log likelihood = 218.7, Nagelkerke $R^2 = .152$; Analysis 3: $\chi^2 = 13.23$, df = 9, $P < .152$; −2 log likelihood = 226.5, Nagelkerke $R^2 = .098$; Analysis 4: $\chi^2 = 27.64$, df = 9, $P < .001$; −2 log likelihood = 212.1, Nagelkerke $R^2 = .196$

[b] This was the only variable block in any analysis statistically significant at $P < .05$ as an omnibus test

analysis. Block 1 is an uncontrolled test of program effectiveness showing that PACT had a significantly higher overall employment rate. Block 2, which tests the predictive power of each client measure, was not statistically significant as an omnibus test in any analysis, in spite of the relatively strong tendency for psychiatric symptoms to discourage work.

Block 3 statistics illustrate the relative sensitivity of the four analyses for detecting variations in outcomes within programs. The age-by-health-by-program interaction term is significant in Analyses 2, 3, and 4, but this block and the full regression model are both significant as omnibus tests only in Analysis 4. Of the four methods, the comparison of cluster analysis-based subgroups provides the statistically strongest evidence of differential program effectiveness.

Specificity

Analysis 4 also provides the greatest specificity as a statistical test of the percentage differences presented in Table 2: older, unhealthy clients were more likely to work than very physically ill, very psychiatrically ill, and relatively healthy clients if they were assigned to PACT, but less likely to work than very physically ill, very psychiatrically ill, or relatively healthy clients if assigned to the clubhouse. Had this same block been significant in Analysis 3, we could conclude that older, physically ill clients had higher work rates than young, psychiatrically well clients in PACT (67% vs. 39%), but the opposite was true for the clubhouse (22% vs. 53%). Findings for this same block in Analysis 2 would be interpreted simply as better work rates for older, unhealthy PACT clients in comparison to everyone else in the study.

One reason that Analysis 4 was the most sensitive and specific test of program differential effectiveness is that cluster analysis is designed to maximize subgroup differences in naturally co-occurring characteristics. As a result, the cluster analysis-based subgroups were distinct in several meaningful ways that aid in the interpretation of findings (Table 3). In addition to having favorable scores on all four

Table 2 Employment rates for sample subgroups within PACT and clubhouse programs

Cluster-based subgroups	Full sample (N = 174)			Work interest sample (N = 121)		
	PACT (N = 85) n (%)	Clubhouse (N = 89) n (%)	Totals N (%)	PACT(n = 63) n (%)	Clubhouse (n = 58) n (%)	Totals N (%)
Very psychiatrically ill	6 (29)	4 (29)	10 (29)	6 (40)	3 (33)	9 (38)
Substance use disorder	12 (67)	3 (23)	15 (48)	12 (**80**)	2 (25)	14 (61)
Very physically ill	3 (33)	9 (41)	12 (39)	3 (50)	9 (45)	12 (46)
Older, unhealthy	10 (83)	2 (15)	12 (48)	8 (**89**)	2 (29)	10 (63)
Relatively healthy	16 (64)	14 (52)	30 (58)	11 (61)	11 (**79**)	22 (69)
Totals	47 (55)	32 (36)	79 (45)	40 (64)	27 (47)	67 (55)

clustering variables, the *relatively healthy subgroup* had better work histories and reported the fewest limitations to everyday activity. The *very psychiatrically ill subgroup* had the highest percentage of schizophrenia spectrum diagnoses and fewest substance use disorders. Individuals in the *very physically ill subgroup* scored highest on the physical health severity measure and reported the most physical limitations to everyday activity. Individuals in the *older, chronically physically ill subgroup* were older when first hospitalized and had the highest self-esteem, but they were the most obese and least likely to have worked in 5 years. All individuals in the *substance use disorder subgroup* had lifestyles of recurrent, disruptive substance abuse, and so were the most frequently homeless or incarcerated. Although each of the cluster analysis-derived subgroups differed ($P < .05$) from the other four subgroups in these defining ways, the five subgroups were comparable in gender, ethnicity, referral source, and program preference at time of project enrollment. As Table 2 shows, the overall pattern of subgroup differences in work rates remains stable even when the sample is reduced to individuals interested in work at the time they were randomized to the PACT or clubhouse program.

Similar subgroups could be created using various cut-points on the four key variables, but a search for optimal groupings would increase Type I errors. The cluster analysis approach required subjective judgment in the selection of an optimal cluster solution, but was guided by findings from Rubin and Panzano (2002), who identified five similar clusters for a sample of 3,600 adults with serious mental illness. Our replication of their groupings with a smaller sample and different variable measures suggests these five groupings are robust and representative of the population.

Aim II. Tests of the Internal Validity of Main Study Findings

To check on the validity of our main study findings, we repeated Analysis 4 (Table 1) with a preceding block of two attitudinal variables (program preference, work interest) known to predict work outcomes (Macias et al. 2005; Macias et al. 2001). We changed the reference group for this formal outcome analysis to the relatively healthy subgroup, since this would be a logical comparison group to most providers and service planners.

As can be seen in Table 4, the program variable (Block 2) is again significant, even when controlling for the attitudinal variables in Block 1. However, the addition of the fourth block of program-by-subgroup interaction terms shows that the PACT and clubhouse work rates (Table 2) differ significantly for two clusters: PACT was more effective for older individuals who had chronic physical health problems than for relatively healthy clients, while the reverse was true within the clubhouse. With the addition of interaction terms, the beta for program assignment is no longer statistically significant ($P = .894$), and is substantially reduced ($\beta = -.09$, SE = .63), indicating that the strong effectiveness of PACT for older adults with health problems accounts for the overall significantly higher work rates for PACT versus clubhouse (variable change not in table).

When total job search hours (log-transformed) is added to the regression model (not shown in table), this dosage variable predicts employment over and above all other variables ($\beta = .28$, SE = .08, $P = .001$), but does not account for program differences in effectiveness because the beta for the program-by-older, chronically physically ill subgroup interaction term remains significant with no decrease in value.

Discussion

The inclusion of a final block of program-by-subgroup interaction terms in our regression analysis of employment rates (Table 4) qualifies the main finding of higher work rates for PACT versus clubhouse, restricting PACT greater vocational effectiveness to a portion of the study sample. The 83% employment rate (Table 2) attained by PACT for clients with age-related chronic health problems not only

Table 3 Characteristics of participants included in the five cluster analysis-based subgroups ($N = 174$)

	Severely mentally ill ($N = 35$)			Severely physically ill ($N = 31$)			Chronically physically ill ($N = 25$)			Substance use disorder ($N = 31$)			Relatively healthy ($N = 52$)		
	$M \pm$ SD	N	%	$M \pm$ SD	N	%	$M \pm$ SD	N	%	$M \pm$ SD	N	%	$M \pm$ SD	N	%
Clustering variables[a]															
Psychiatric symptom severity	**56.46 ± 6.33**[b]			44.97 ± 5.98			**39.26 ± 5.72**[b]			48.68 ± 6.96			**41.37 ± 5.20**[b]		
Physical illness severity	**13.67 ± 11.89**[b]			**66.59 ± 20.23**[b]			21.59 ± 10.67			**12.28 ± 12.27** [b]			**8.29 ± 9.61**[b]		
Substance use severity	**.01 ± .28**[b]			.74 ± .89			**.28 ± .54**[b]			**2.00 ± .00**[b]			**.15 ± .36**[b]		
Baseline age in years	**34.74 ± 10.16**[b]			40.71 ± 8.48			**52.20 ± 7.21**[b]			35.97 ± 6.22			**33.19 ± 7.60**[b]		
Comparison variables															
Female gender		15	43		16	52		15	60		13	42		20	39
Minority race or ethnicity		10	29		3	10		3	12		8	26		13	25
Schizophrenia diagnosis		**25**	**71**[b]		14	45		9	36		17	55		25	48
Worked in past five years		22	63		14	45		**9**	**36**[b]		19	61		**36**	**69**[b]
Homeless at any time		2	6		4	13		3	12		**9**	**29**[b]		4	8
Incarcerated at any time		**1**	**3**[b]		4	13		3	12		**12**	**39**[b]		6	12
Self-esteem[c]	26.09 ± 4.84			26.73 ± 5.11			**29.56 ± 5.53**[b]			27.29 ± 4.42			27.83 ± 5.66		
Obesity[d]		4	11		4	13		**8**	**32**[b]		5	16		9	17
Physical role limitations[e]	2.00 ± 1.41			**2.33 ± 1.17**[b]			1.87 ± 1.41			1.78 ± 1.42			**1.45 ± 1.30**[b]		
Age at first hospitalization	25.54 ± 9.13			25.16 ± 8.36			**36.32 ± 14.04**[b]			26.94 ± 8.97			25.77 ± 8.34		
Experimental variables															
Assertive community treatment		21	25		**9**	**11**[b]		12	14		18	21		25	29
Clubhouse		14	16		**22**	**25**		13	15		13	15		27	30
Assigned non-preferred program		13	37		6	19		7	28		6	19		15	29
Baseline interest in work		24	69		26	84		16	64		23	74		32	62

[a] The omnibus ANOVA for each clustering variable (df = 4, 169) was significant at $P < .001$

[b] This subgroup differed significantly from the other 4 subgroups combined at $P < .05$ within a chi square or logistic regression analysis

[c] Total score during project on Rosenberg's Self-Esteem Scale. Scores range from 10 to 40, with higher scores indicating higher self-esteem

[d] Medicaid claims data treatment for obesity or interviewer report of severe obesity (1) versus a lack of this evidence of obesity (0)

[e] Mean score during project on role limitations subscale of MOS SF-36. Scores range from 0 to 4, with higher scores indicating lower functioning

Table 4 Logistic regression analyses of work rates for cluster analysis-based subgroups assigned to PACT and clubhouse programs ($N = 174$)[a]

Predictor variables	β (log odds)	SE	P	Exp(β) (odds ratio)	*Wald*
Block 1: control variables[b]					
Match to service preference	−.19	.40	.627	.82	.24
Mismatch to service preference	−.52	.39	.180	.59	1.80
Work interest	1.43	.38	.001	4.18	14.33
Block 2: experimental condition[c]					
Program assignment (PACT)	.70	.33	.035	2.01	4.46
Block 3: participant subgroups[d]					
Very psychiatrically ill	−1.62	.52	.002	.20	9.68
Substance use disorder	−.81	.52	.116	.44	2.48
Very physically ill	−1.13	.52	.029	.32	4.76
Chronically physically ill	−.54	.55	.323	.58	.98
Block 4: subgroups within programs[e]					
Program × very psychiatrically ill	−.11	1.03	.912	.89	.01
Program × substance use disorder	1.60	1.07	.135	4.94	2.23
Program × very physically ill	−.21	1.07	.843	.81	.04
Program × chronically physically ill	3.31	1.33	.013	27.36	6.22

[a] Full regression model: $\chi^2 = 45.84$, df = 12, $P < .001$; fit statistics: −2 log likelihood = 193.9, Nagelkerke $R^2 = .310$

[b] Block 1: $\chi^2 = 18.52$, df = 3, $P < .001$. Match/Mismatch = randomized to preferred versus not preferred program; reference category was no preference

[c] Improvement in model with addition of program assignment: $\chi^2 = 4.51$, df = 1, $P = .034$. Assertive community treatment (PACT), 1; clubhouse, 0

[d] Improvement in model with addition of Block 3: $\chi^2 = 11.93$, df = 4, $P = .018$

Each of these four subgroups was compared to the relatively healthy subgroup

[e] Improvement in model with addition of Block 4: $\chi^2 = 10.88$, df = 4, $P = .028$

surpassed the clubhouse work rates for this subgroup, but also the overall 50–55% benchmark employment rates reported for specialized supported employment teams that usually screen for work interest (Cook et al. 2005; Macias et al. 2006; Twamley et al. 2003). This PACT rate rises to 89% when we consider only those clients who expressed an interest in work at the time they enrolled in the study ($N = 121$). These findings should be useful for service planning because older adults with chronic health problems are now prevalent in the population of adults with severe mental illness (Jones et al. 2004; Rubin and Panzano 2002) and are likely to increase in number as the baby-boom generation grows older.

Advantages of Cluster Analysis

Use of cluster analysis for subgroup identification not only increased statistical power, but also solved the dilemma of how to blend the advantages of sample homogeneity and heterogeneity. Sample subgroups differed greatly from one another, but each was homogeneous in the sense that individuals within that subgroup shared the same mix of correlated characteristics, regardless of how heterogeneous the particular mix. This balance of heterogeneity and homogeneity allowed rich, complex descriptions of differential program effectiveness. Using variables, we can report only that PACT became more vocationally effective as client age increased, assuming all other client variables are held constant. Using subgroups, we can say that PACT was especially effective for middle-aged and older clients with chronic health problems who became psychiatrically ill later in life.

Study Limitations

Our subgroups were identified after randomization, so our findings require replication in new studies that first identify subgroups and then randomly assign individuals within subgroups to experimental conditions. Hopefully, our subgroup descriptions will also prove useful in the design of stratified sampling to ensure comparably sized subgroups. New studies are needed not only to test the reliability of these particular findings, but also to test more specific hypotheses, such as whether a greater appreciation of staff outreach and monitoring by older, chronically physically ill adults might account for their better employment rates in PACT, or whether older, chronically physically ill adults assigned to the clubhouse found the

option of voluntary clubhouse work more appealing than a job in the competitive workforce.

A caution is also warranted: while our cluster definitions may be useful in the design of new studies, they should not become standardized subgroup definitions for the population of adults with severe mental illness. Population groupings found to predict outcomes in one study may not be as meaningful when used to examine other programs or other outcomes. Hypotheses should always be outcome and program specific. Moreover, even if studies have similar designs, target similar populations, and test similar interventions, conceptual replications that measure the same concepts in different ways are more useful than literal replications for the refinement of program theory (Aronson et al. 1990).

Conclusions

Only a handful of service evaluations have disaggregated heterogeneous samples into homogeneous subgroups to test hypotheses about the relative effectiveness of particular interventions for certain types of individuals, and these few studies span several service fields (Abel et al. 2005; Carey et al. 2007; Clark and Rich 2003; Halvorsen and Monsen 2007; Hodges et al. 2004; Maisto et al. 2001; McKendrick et al. 2007; Ogrodniczuk et al. 2007). We hope that these exemplary studies and our own subgroup-based findings will encourage the formulation and testing of theory-based hypotheses about differential service effectiveness. We also hope that the advantages of subgroup analyses for ensuring valid interpretations of whole-sample outcomes will encourage the publication of null findings (Turner et al. 2008) and spur agencies that fund services research, and registries for randomized controlled trials, to routinely require a priori hypotheses that predict the relative effectiveness of experimental interventions for particular sample subgroups.

Acknowledgments This project was funded by an interdisciplinary research grant from the National Institute of Mental Health (R01-MH-62628), and was part of the multi-project Employment Intervention Demonstration Program (EIDP) funded by the Substance Abuse and Mental Health Service Administration (SM-51831). Supplemental support was provided by the van Ameringen Foundation, the John D. and Catherine T. MacArthur Foundation, Massachusetts Department of Mental Health, and the University of Massachusetts Medical School. No author, affiliated institution, or funding organization has a conflict of interest. The authors thank their NIMH project monitor, Ann Hohmann, Ph.D., for her helpful insights and encouragement.

References

Abel, K. M., Webb, R. T., Salmon, M. P., Wan, M. W., & Appleby, L. (2005). Prevalence and predictors of parenting outcomes in a cohort of mothers with schizophrenia admitted for joint mother and baby psychiatric care in England. *Journal of Clinical Psychiatry, 66*(6), 781–789.

Ackerman, R. J. (1999). An interactional approach for pharmacopsychologists and psychologists: Gender concerns. *Journal of Clinical Psychology in Medical Settings, 6*(1), 39–61.

Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York: Guilford Press.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82*(1), 192–201.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Allness, D., & Knoedler, W. (1998). *The PACT model: A Manual for PACT Start-Up*. Arlington, VA: National Alliance for the Mentally Ill.

Anderson, S. B. (1998). *We are not alone: Fountain house and the development of clubhouse culture*. New York, NY: Fountain House, Inc.

Anthony, W. A., Rogers, E. S., Cohen, M., & Davies, R. R. (1995). Relationships between psychiatric symptomatology, work skills, and future vocational performance. *Psychiatric Services, 46*(4), 353–358.

Aronson, E., Ellsworth, P., Carlsmith, J. M., & Gonzales, M. H. (1990). *Methods of research in social psychology*. New York: McGraw-Hill.

Batstra, L., Bos, E. H., & Neeleman, J. (2002). Quantifying psychiatric comorbidity: Lessons from chronic disease epidemiology. *Social Psychiatry and Psychiatric Epidemiology, 37*, 105–111.

Beard, J. H., Propst, R. N., & Malamud, T. J. (1982). The fountain house model of psychiatric rehabilitation. *Psychosocial Rehabilitation Journal, 5*(1), 1–12.

Bekker, M. H. J. (2003). Investigating gender within health research is more than sex disaggregation of data: A multi-facet gender and health model. *Psychology, Health and Medicine, 8*(2), 231–243.

Beutler, L., Brown, M., Crothers, L., Booker, K., & Seabrook, M. (1996). The dilemma of factitious demographic distinctions in psychological research. *Journal of Consulting and Clinical Psychology, 64*, 892–902.

Bickman, L., Noser, K., & Summerfelt, W. T. (1999). Long-term effects of a system of care on children and adolescents. *Journal of Behavioral Health Services and Research, 26*(2), 185–202.

Blankertz, L. (1998). The value and practicality of deliberate sampling for heterogeneity: A critical multiplist perspective. *American JoumaI of Evaluation, 19*(3), 307–324.

Bond, G. R., Vogler, K., Resnick, S. G., Evans, L., Drake, R., & Becker, D. (2001). Dimensions of supported employment: Factor structure of the IPS fidelity scale. *Journal of Mental Health (UK), 10*(4), 383–393.

Bühringer, G. (2006). Allocating treatment options to patient profiles: Clinical art or science? *Addiction, 101*, 646–652.

Burke-Miller, J., Cook, J. A., Grey, D. D., Razzano, L., Blyler, C. R., Leff, H. S., et al. (2006). Demographic characteristics and employment among people with severe mental illness in a multisite study. *Community Mental Health Journal, 42*(2), 143–159.

Carey, K. B., Henson, J. M., Carey, M. P., & Maisto, S. A. (2007). Which heavy drinking college students benefit from a brief motivational intervention? *Journal of Consulting and Clinical Psychology, 75*(4), 663–669.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*(1), 7–18.

Chwastiak, L. A., Rosenheck, R. A., McEvoy, J. P., Keefe, R. S. E., Swartz, M. S., & Lieberman, J. A. (2006). Interrelationships of psychiatric symptom severity, medical comorbidity, and

functioning in schizophrenia. *Psychiatric Services, 57*(8), 1102–1109.

Clark, C., & Rich, A. R. (2003). Outcomes of homeless adults with mental illness in a housing program and in case management only. *Psychiatric Services, 54*(1), 78–83.

Cook, J. A., Pickett-Schenk, S. A., Gray, D., Banghart, M., Rosenheck, R., & Randolph, F. (2001). Vocational outcomes among formerly homeless persons with severe mental illness in the ACCESS program. *Psychiatric Services, 52*(8), 1075–1080.

Cook, J. A., Leff, H. S., Blyler, C. R., Gold, P. B., Goldberg, R. W., Mueser, K. T., et al. (2005). Results of a multisite randomized trial of supported employment interventions for individuals with severe mental illness. *Archives of General Psychiatry, 62*, 505–512.

Dembling, B. P., Chen, D. T., & Vachon, L. (1999). Life expectancy and causes of death in a population treated for serious mental illness. *Psychiatric Services, 50*(8), 1036–1042.

Department of Labor. (1998). Job Training Program Act, Disability Grant Program Funded Under Title III, Section 323, and Title IV, Part D, Section 452 (Vol. 63): Federal Register.

Dickey, B., Normand, S.-L., Weiss, R., Drake, R., & Azeni, H. (2002). Medical morbidity, mental illness, and substance use disorders. *Psychiatric Services, 53*(7), 861–867.

Dixon, L., Goldberg, R., Lehman, A., & McNary, S. (2001). The impact of health status on work, symptoms, and functional outcomes in severe mental illness. *Journal of Nervous and Mental Disease, 189*, 17–23.

Drebing, C., Van Ormer, A., Krebs, C., Rosenheck, R., Rounsaville, B., Herz, L., et al. (2005). The impact of enhanced incentives on vocational rehabilitation outcomes for dually diagnosed veterans. *Journal of Applied Behavior Analysis, 38*(3), 359–372.

Druss, B. G., Marcus, S. C., Rosenheck, R. A., Olfson, M., Tanielian, T., & Pincus, H. A. (2000). Understanding disability in mental and general medical conditions. *American Journal of Psychiatry, 157*(9), 1485–1491.

Frey, J. L. (1994). Long term support: The critical element to sustaining competitive employment: Where do we begin? *Psychosocial Rehabilitation Journal, 17*(3), 127–133.

Goldberg, R. W., Lucksted, A., McNary, S., Gold, J. M., Dixon, L., & Lehman, A. (2001). Correlates of long-term unemployment among inner-city adults with serious and persistent mental illness. *Psychiatric Services, 52*(1), 101–103.

Halvorsen, M. S., & Monsen, J. T. (2007). Self-image as a moderator of change in psychotherapy. *Psychotherapy Research, 17*(2), 205–217.

Hendryx, M. S., Moore, R., Leeper, T., Reynolds, M., & Davis, S. (2001). An examination of methods for risk-adjustment of rehospitalization rates. *Mental Health Services Research, 3*(1), 15–24.

Hendryx, M. S., & Teague, G. B. (2001). Comparing alternative risk-adjustment models. *Journal of Behavioral Health Services and Research, 28*(3), 247–257.

Hodges, K., Xue, Y., & Wotring, J. (2004). Outcomes for children with problematic behavior in school and at home served by public mental health. *Journal of Emotional and Behavioral Disorders, 12*(2), 109–119.

Howell, W., & Peterson, P. (2004). Uses of theory in randomized field trials. *American Behavioral Scientist, 47*(5), 634–657.

James, G. M., Sugar, C. A., Desai, R., & Rosenheck, R. A. (2006). A comparison of outcomes among patients with schizophrenia in two mental health systems: A health state approach. *Schizophrenia Research, 86*, 309–320.

Jensen, P., Arnold, E., Richters, J., Severe, J., Vereen, D., Vittiello, B., & Shiller, E. (1999). Moderators and mediators of treatment response for children with attention-deficit/hyperactivity disorder. *Archives of General Psychiatry, 56*, 1088–1096.

Jones, D. R., Macias, C., Barreira, P. J., Fisher, W. H., Hargreaves, W. A., & Harding, C. M. (2004). Prevalence, severity, and co-occurrence of chronic physical health problems of persons with serious mental illness. *Psychiatric Services, 55*(11), 1250–1257.

Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. New York: Cambridge University Press.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin, 13*, 261–276.

Kenny, D. A., Calsyn, R. J., Morse, G. A., Klinkenberg, W. D., Winter, J. P., & Trusty, M. L. (2004). Evaluation of treatment programs for persons with severe mental illness: Moderator and mediator effects. *Evaluation Review, 28*(4), 294–324.

King, R. D., Gaines, L. S., Lambert, E. W., Summerfelt, W. T., & Bickman, L. (2000). The co-occurrence of psychiatric substance use diagnoses in adolescents in different service systems: Frequency, recognition, cost, and outcomes. *Journal of Behavioral Health Services and Research, 27*(4), 417–430.

Kraemer, H. C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry, 158*(6), 848–856.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry, 59*, 877–883.

Kraemer, H. C., Wilson, K. A., & Hayward, C. (2006). Lifetime prevalence and pseudocomorbidity in psychiatric research. *Archives of General Psychiatry, 63*, 604–608.

Krause, M. S., & Howard, K. I. (2003). What random assignment does and does not do. *Journal of Clinical Psychology, 59*, 751–766.

Kronick, R., Gilmer, T., Dreyfus, T., & Lee, L. (2000). Improving health-based payment for medicaid beneficiaries: CDPS. *Health Care Financing Review, 21*(3), 29–64.

Lehman, A. F., Goldberg, R. W., Dixon, L. B., NcNary, S., Postrado, L., Hackman, A., et al. (2002). Improving employment outcomes for persons with severe mental illnesses. *Archives of General Psychiatry, 59*(2), 165–172.

Lipchik, G. L., Nicholson, R. A., & Penzien, D. B. (2005). Allocation of patients to conditions in headache clinical trials: Randomization, stratification, and treatment matching. *Headache, 45*, 419–428.

Macias, C., DeCarlo, L., Wang, Q., Frey, J., & Barreira, P. (2001). Work interest as a predictor of competitive employment: Policy implications for psychiatric rehabilitation. *Administration and Policy in Mental Health, 28*(4), 279–297.

Macias, C., Barreira, P., Hargreaves, W., Bickman, L., Fisher, W. H., & Aronson, E. (2005). Impact of referral source and study applicants' preference for randomly assigned service on research enrollment, service engagement, and evaluative outcomes. *American Journal of Psychiatry, 162*(4), 781–787.

Macias, C., Rodican, C. F., Hargreaves, W. A., Jones, D. R., Barreira, P. J., & Wang, Q. (2006). Supported employment outcomes of a randomized controlled trial of assertive community treatment and clubhouse models. *Psychiatric Services, 57*(10), 1406–1415.

Maisto, S. A., Conigliaro, J., McNeil, M., Kraemer, K., & Kelley, M. E. (2001). The relationship between eligibility criteria for participation in alcohol brief intervention trials and other alcohol and health-related variables. *American Journal of Addictions, 10*(3), 218–231.

McBride, C., Atkinson, L., Quilty, L. C., & Bagby, R. M. (2006). Attachment as moderator of treatment outcome in major depression: A randomized control trial of interpersonal psychotherapy versus cognitive behavior therapy. *Journal of Consulting and Clinical Psychology, 74*(6), 1041–1054.

McKendrick, K., Sullivan, C., Banks, S., & Sacks, S. (2007). Modified therapeutic community treatment for offenders with MICA disorders: Antisocial personality disorder and treatment outcomes. *Journal of Offender Rehabilitation, 44*(2–3), 133–159.

Miklowitz, D., & Clarkin, J. F. (1999). Balancing internal and external validity. *Prevention and Treatment, 2*, 1–4.

Mueser, K. T., Salyers, M. P., & Mueser, P. R. (2001). A prospective analysis of work in schizophrenia. *Schizophrenia Bulletin, 27*(2), 281–296.

Ogrodniczuk, J. S., Joyce, A. S., & Piper, W. E. (2007). Effect of patient dissatisfaction with the therapist on group therapy outcome. *Clinical Psychology and Psychotherapy, 14*(2), 126–134.

Peck, L. R. (2005). Using cluster analysis in program evaluation. *Evaluation Review, 29*(2), 178–196.

Pettinati, H. M., Volpicelli, J. R., Kranzler, H. R., Luck, G., Rukstalis, M. R., & Cnaan, A. (2000). Sertraline treatment for alcohol dependence: Interactive effects of medication and alcoholic subtype. *Alcoholism: Clinical and Experimental Research, 24*(7), 1041–1049.

Propst, R. (1992). Standards for clubhouse programs: Why and how they were developed. *Psychosocial Rehabilitation Journal, 16*(2), 25–30.

Rapkin, B. D., & Dumont, K. A. (2000). Methods for identifying and assessing groups in health behavioral research. *Addiction, 95*(Supplement 3), S395–S417.

Razzano, L., & Hamilton, M. M. (2005). Health-related barriers to employment among people with HIV/AIDS. *Journal of Vocational Rehabilitation, 22*(3), 179–188.

Razzano, L. A., Cook, J. A., Burke, J., Mueser, K. T., Pickett-Schenk, S., Grey, D. D., et al. (2005). Clinical factors associated with employment among people with severe mental illness: Findings from the employment intervention demonstration program. *Journal of Nervous and Mental Disease, 193*(11), 705–713.

Regenold, M., Sherman, M., & Fenzel, M. (1999). Getting back to work: Self-efficacy as a predictor of employment outcome. *Psychiatric Rehabilitation Journal, 22*(4), 361–367.

Rothwell, P. M. (2005). External validity of randomized controlled trials: ''To whom do the results of this trial apply?''. *Lancet, 365*, 82–93.

Rubin, W. V., & Panzano, P. C. (2002). Identifying meaningful subgroups of adults with severe mental illness. *Psychiatric Services, 53*(4), 452–457.

Ruscio, A. M., & Holohan, D. R. (2006). Applying empirically supported treatments to complex cases: Ethical, empirical, and practical considerations. *Clinical Psychology: Science and Practice, 13*(2), 146–162.

Salyers, M. P., McHugo, G. H., Cook, J. A., Razzano, L. A., Drake, R. E., & Mueser, K. T. (2001). Reliability of instruments in a cooperative, multisite study: Employment intervention demonstration program. *Mental Health Services Research, 3*(3), 129–139.

Slade, E., & Salkever, D. (2001). Symptom effects on employment in a structural model of mental illness and treatment: Analysis of patients with schizophrenia. *Journal of Mental Health Policy and Economics, 4*(1), 25–34.

SPSS. (1999). Statistical Program for the Social Sciences (Version 11) [Windows]. Chicago.

Stein, L. I., & Test, M. A. (1980). Alternative to mental hospital treatment. I. Conceptual model, treatment program, and clinical evaluation. *Archives of General Psychiatry, 37*(4), 392–397.

Stone-Romero, E. F., & Anderson, L. E. (1994). Relative power of moderated multiple regression and the comparison of subgroup correlation coefficients for detecting moderating effects. *Journal of Applied Psychology, 79*(3), 354–359.

Stout, C. E., & Hayes, R. A. (Eds.). (2005). *Evidence-based practice: Methods, models, and tools for mental health professionals*. New York, NY: John Wiley.

Trach, J. S. (1990). Supported employment program characteristics. In R. Rusch (Ed.), *Supported employment: Models, methods and issues*. Sycamore, IL: Sycamore Press.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*, 252–260.

Twamley, E., Jeste, D. V., & Lehman, A. (2003). Vocational rehabilitation in schizophrenia and other psychotic disorders: A literature review and meta-analysis of randomized controlled trials. *Journal of Nervous and Mental Disease, 191*(8), 515–523.

Uehara, E., Srebnik, D., & Smukler, M. (2003). Statistical and consensus-based strategies for grouping consumers in mental health level-of-care schemes. *Administration and Policy in Mental Health, 30*(4), 287–306.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.

Wells, K. B. (1999). Treatment research at the crossroads: The scientific interface of clinical trials and effectiveness research. *American Journal of Psychiatry, 156*(1), 5–10.

West, S. G., & Aiken, L. S. (1997). Toward understanding individual effects in multicomponent prevention programs: Design and analysis strategies. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research*. Washington, DC: American Psychological Association.

Wewiorski, N. J., & Fabian, E. S. (2004). Association between demographic and diagnostic factors and employment outcomes for people with psychiatric disabilities: A synthesis of recent research. *Mental Health Services Research, 6*(1), 9–21.

Workforce Investment Act of (1998): Public law 105–220.