Article

Stability of Cognitive Performance in Older Patients With Schizophrenia: An 8-Week Test-Retest Study

Philip D. Harvey, Ph.D.

Barton W. Palmer, Ph.D.

Robert K. Heaton, Ph.D.

Somaia Mohamed, M.D., Ph.D.

John Kennedy, M.D.

Adam Brickman, M.A.

Objective: It is important to understand the stability of cognitive performance in schizophrenia in order to understand 1) the potential improvements in performance associated with the beneficial effects of cognitive-enhancing medications and 2) the potential variation in performance on cognitive measures that are components of neuroimaging studies. There are several factors that could lead to spurious improvements in cognitive test scores, including practice effects and random test-retest errors.

Method: Forty-five older patients with schizophrenia who were receiving conventional antipsychotic medications participated in the study. All subjects completed a comprehensive neuropsychological test battery at baseline and again at an 8-week follow-up evaluation.

Results: Performance on all of the cognitive measures was stable over time, as evidenced by significant test-retest correlations, and practice effects were generally absent or minimal. Tests administered with alternate forms were no more temporally stable than tests administered twice with the same form. Very few individual cases had substantial variation at retest across the 22 test scores. From these data, "norms for evaluating change" were developed and described using the reliable change index method.

Conclusions: The data suggest that the findings of modest cognitive improvements seen in some prior studies of schizophrenia patients when treated with second-generation antipsychotic medications were probably not due to simple retesting artifacts. At the same time, because of the variance in some of these test-retest performances, relatively substantial changes in performance on the part of individual patients would be required to be clearly interpretable.

(Am J Psychiatry 2005; 162:110–117)

Previous research has focused on cognitive enhancement in schizophrenia (1). An extensive literature has developed on the purported cognitive-enhancing effects of second-generation (a.k.a. "atypical") antipsychotic medications (2), and in addition there has recently been increased interest in identification of compounds chosen specifically for their cognitive-enhancing potential (3, 4). These studies have examined cognitive enhancement effects on a wide array of cognitive ability areas, including episodic memory (5), attention/vigilance (6), executive functioning (7), psychomotor speed (8), and verbal skills (9). While some of these studies have used adequate clinical trial methodology, the majority have had substantial methodological limitations. For instance, many of the studies did not use blinded designs or random assignment to treatment conditions (see Harvey and Keefe [1] for a review of these issues). There has been equivalent interest in developing cognitive-enhancing techniques that are behavioral in nature, with many of the same issues applying to these studies as well.

One of the most salient methodologic issues that must be considered in interpreting apparent improvements in cognitive test performance is whether such changes actually reflect improved cognitive ability or simply result from improved test performance without improvement in the underlying cognitive ability. Test performance may appear to improve because of the combined influence of normal variability in test performance (due to measurement error) plus examinees' increased familiarity with the test materials and procedures over repeated exposure ("practice effects"). This issue also applies to nonpharmacological interventions (i.e., cognitive remediation) and to inferences regarding the importance of correlations between cognitive test performance and the results of neuroimaging studies. Unreliable measurement would reduce the validity of inferences associated with correlations between test performance and imaging results, but few data are available to address the stability of performance on these cognitive measures.

One method of addressing practice effects is the use of alternate forms, yet full equivalence of forms is an ideal that is difficult to achieve, and many of the cognitive tests that are most widely used in the schizophrenia literature have only one version. Moreover, alternate forms do not fully eliminate practice effects. There appear to be at least two potential components to practice effects: 1) an examinee's performance may improve with repeated testing because he or she learns the specific item content over re-

peated presentations, and 2) an examinee's performance may improve with repeated testing because he or she becomes more familiar with performing tasks similar to the target assessment battery (method variance) or becomes more familiar or comfortable with being administered neurocognitive tests in general. Parallel forms are rarely perfectly parallel, but even when ideal they correct for effects listed under the first component but do not correct for those subsumed under the second.

To some degree, interpretive difficulties in clinical trials that are due to measurement error and practice effects can be circumvented by study designs that include double-blind random assignment of some participants to a placebo (or other noncognitive enhancing) arm. However, there have been increasingly frequent concerns expressed about the use of placebo-controlled designs in schizophrenia research. Also, even with an adequately controlled design, while one may be able to show differences in group mean performance, it is difficult to evaluate whether changes have occurred on the level of an individual patient. This level of change is truly the endpoint goal of cognitive enhancement treatments, regardless of mechanism.

Identification of meaningful changes in performance on the part of schizophrenia patients, either due to improvements in performance associated with treatment or deterioration in clinical status, requires understanding of normative patterns of test-retest stability (10, 11). Test-retest data from healthy comparison subjects may provide some insight, but norms from cognitively stable schizophrenia patients are needed because patients may show more test-retest error variance than healthy individuals (12). With the advent of newer antipsychotic medications with reduced severity of side effects (13) and greater efficacy in critical symptom domains (14), it is difficult to perform clinical trials in which patients have the possibility of being randomly assigned to receive older antipsychotic treatments. If normative data were collected on clinically stable individuals receiving older medications, these data might be helpful in evaluating changes seen in future studies examining compounds with purported cognitive-enhancing effects. The purpose of the present study was threefold: 1) to examine practice effects over a test-retest interval similar to a clinical trial, 2) to develop norms for evaluating cognitive change among middle-aged and older psychotic patients treated with conventional antipsychotic medications, and 3) to identify the width of the prediction intervals for defining "unusual" changes that would be unlikely to be due to either measurement error or practice effects. We also examined whether retest stability and prediction intervals varied as a function of baseline levels of cognitive impairment, in that patients with schizophrenia tend to vary in their levels of impairment and it may provide useful information for later trials in terms of differential variance associated with greater or lesser levels of impairment.

This study used a neuropsychological battery (the Aged Schizophrenia Assessment Schedule–Cognitive) that was originally designed by an advisory team of neuropsychological experts for Eli Lilly and Company for use in multisite studies of the cognitive effects of antipsychotic medications among middle-aged and older patients with psychotic disorders. In the present study, we provide a comprehensive study of retest effects on cognitive functioning in schizophrenia, performed on 45 clinically and cognitively stable outpatients who were receiving stable doses of conventional antipsychotic medications. Participants completed the Aged Schizophrenia Assessment Schedule–Cognitive battery at baseline and at an 8-week follow-up evaluation during a period of clinical stability without changes in their treatment status.

Method

Subjects

Participants were 45 middle-aged and older patients with schizophrenia (N=33) or schizoaffective disorder (N=12) who were recruited and evaluated at any of three research sites, the University of California at San Diego Department of Psychiatry, Mt. Sinai School of Medicine, or the University of Cincinnati Department of Psychiatry. All participants were 45 years of age or older, each was receiving clinically appropriate doses of conventional neuroleptic medications during the study, and there had been no changes in their medication treatment during the month preceding the baseline evaluation. We excluded patients with a history of other psychiatric diagnoses or medical conditions that might impact their cognitive functioning (e.g., substance abuse or dependence, neurologic conditions, head trauma with loss of consciousness, seizure disorder, or degenerative dementia). Patients were also excluded if their current antipsychotic medication dose was greater than 1500 mg/day in chlorpromazine equivalents, they were receiving selective serotonin reuptake inhibitor antidepressants, or they had a current diagnosis of major depression. These criteria were intended to ensure that the patients in the sample were clinically stable over the test-retest interval.

Measures

Psychopathology and motor symptoms. Level and changes in severity of psychopathology were assessed with the positive and negative subscale scores of the Positive and Negative Syndrome Scale (15) and the Montgomery-Åsberg Depression Rating Scale (16). Severity of extrapyramidal symptoms was assessed with a modified version of the Simpson-Angus Rating Scale (17), which examined nine items (gait, balance, arm dropping, rigidity of major joints, cogwheeling, glabella tap reflex, tremor, salivation, and akinesia). It had a potential score range of 0 to 30: three items (balance, cogwheeling, glabella tap) were rated from 0 to 2, the other six were rated from 0 to 4. Motor abnormalities were also evaluated with the global clinical assessment of akathisia score from the Barnes Rating Scale for Drug-Induced Akathisia (18). The latter score has a potential score range from 0 (absent) to 5 (severe akathisia).

Neuropsychological battery. The Aged Schizophrenia Assessment Schedule–Cognitive battery consists of the following specific components: total correct on the McGurk Visual Spatial Working Memory Test (19); total learning and delayed recall scores from the Word List Learning Test (20); seconds to complete parts A and B of the Trail Making Test (21); d-prime and omission error totals from the Continuous Performance Test–Identical

Pairs Version (22); total raw score on the digit span subtest, raw score for the letter-number sequencing task, and total correct for digit symbol-coding, all taken from the WAIS-III (23); raw scores for total perseverative errors and total categories completed in the 64-card Wisconsin Card Sorting Test (24); total correct for the letter fluency task; total words for the animal fluency task; right- and left-hand totals for the Finger Tapping Test; raw score (total correct minus total errors) on the WAIS-III symbol search; total correct and correct minus errors of commission on the digit cancellation task; total correct on the Benton Judgment of Line Orientation (25); total correct on the Hooper Visual Organization Test (26); and total errors on the modified (30-item) version of the Hiscock Digit Memory Test (27).

The measures in this battery provide a comprehensive assessment of a number of constructs relevant to schizophrenia, its treatment, and normal aging, including 1) verbal skills (verbal fluency), 2) attention/working memory and vigilance (McGurk Visual-Spatial Working Memory Test, Continuous Performance Test–Identical Pairs Version, digit cancellation test, digit span, letter-number sequencing), 3) verbal learning and memory (Word List Learning Test), 4) psychomotor speed (Trail Making Test–part A, digit symbol-coding, symbol search), 5) abstraction/problem solving/mental flexibility (64-card Wisconsin Card Sorting Test, Trail Making Test–part B), 6) perceptual organizational ability (Hooper Visual Organization Test, Benton Judgment of Line Orientation Test), 7) motor speed (Finger Tapping Test), and 8) effort/motivation (Hiscock Digit Memory Test).

In designing the Aged Schizophrenia Assessment Schedule–Cognitive battery, the advisory team attempted to incorporate tests that targeted cognitive constructs for which preliminary studies had suggested atypical antipsychotic medications might have a beneficial impact (such as verbal fluency, attention/working memory, verbal learning, and abstraction/problem-solving), as well as some cognitive abilities for which there was not an expectation of improvement (perceptual organization, motivation/cooperation). Another primary consideration in selecting the specific tests to include in this battery was that the battery was being designed for repeated administration to older psychotic patients in clinical trials who might not tolerate lengthier assessments. Two tests were used with alternate forms in a fixed order (digit cancellation and the Word List Learning Test) from baseline to the follow-up evaluation to minimize the degree to which practice effects on these particular measures might be observed from patients' learning the item content.

Procedures

All subjects were assessed individually by experienced staff members under the supervision of doctoral-level neuropsychologists at each site. The Aged Schizophrenia Assessment Schedule–Cognitive battery was administered at baseline and again 8 weeks later by the same site-specific staff member. For some patients, the testing sessions were completed over several days (within a 1-week window) to minimize fatigue.

Statistical Analyses

We calculated the mean and standard deviation of each score at baseline and at the 8-week follow-up evaluation. For each score, we calculated the correlation (Pearson's r) between the baseline and 8-week follow-up score as our index of test-retest reliability. Practice effects were measured in terms of the mean difference between follow-up minus baseline scores. Changes in mean scores were evaluated in terms of paired sample t tests. Significance was defined a priori as $p<0.05$ (two-tailed.)

Norms for change were developed for the Aged Schizophrenia Assessment Schedule–Cognitive variables on the basis of mean practice effect and standard error of measurement associated with each of the variables using the reliable change index plus practice effect method (10–12). Specifically, 90% confidence intervals were developed for the standard error of the difference for each test. The standard error of the difference describes the spread of the distribution of neuropsychological change scores that would be expected if no actual change in cognitive abilities had occurred. The standard error of the difference (SE_{diff}) was determined from the standard error of measurement for each test using the following formula:

$$SE_{diff}=[(SD_x^2+SD_y^2)(1-r_{xy})]^{\frac{1}{2}}$$

The standard error of management (SE_m) was, in turn, estimated from the standard deviation of baseline scores (SD_1) and test-retest reliability (r_{xx}) for each test (calculated in terms of Pearson's r between each respective baseline and 8-week follow-up score), as observed within the present sample, using the following formula:

$$SE_m = SD_1\left[(1-r_{xx})^{\frac{1}{2}}\right]$$

The values of the standard deviation of baseline scores and test-retest reliability, as well as the practice effect (the mean difference between follow-up minus baseline scores), were determined from the present sample. Then, a 90% confidence interval for expected retest scores (X_2) was determined by multiplying the SE_{diff} by ±1.64, using the formula:

$$90\%\text{ confidence interval} = (X_1 + \text{mean practice effect}) \pm 1.64\ SE_{diff}.$$

Thus, X_1 represents the baseline score for each subject, and the mean practice effect equals the mean of the change scores (retest score minus baseline) for all subjects in the sample. Using this definition, 90% of the retest scores by chance alone should fall between the lower and upper boundaries (adjusted for practice effects) of this confidence interval. Retest scores above this boundary would be expected to occur less than 5% of the time and would represent statistically significant improvement; scores below this boundary would reflect statistically significant worsening.

Given a test battery of this size, most subjects may be expected to demonstrate "significant change" by random chance (i.e., on average, at least two scores on a 22-test battery would be expected to exceed a 90% confidence interval). Thus, we also applied the confidence intervals for each test to the subjects within this data set to determine the number of tests that he or she had retest scores that were better, worse, or within the expected confidence intervals. This base rate information is intended to provide a basis by which users of these tests can determine the number of tests on which significant positive or negative changes must be observed in order to exceed the level expected by chance alone.

Finally, we examined all of our test-retest differences as a function of the baseline levels of cognitive functioning in the subjects. We expected that patients with varying levels of cognitive impairment might have differences in their error variance, although we did not have a specific hypothesis as to which group would demonstrate more variability. In order to define a subgroup of patients with general cognitive impairment, we created a summary score that was based on performance across 11 scores for which we had sufficient normative basis to define impairment (Trail Making Test parts A and B, digit span, letter-number sequencing, digit symbol, symbol search, 64-card Wisconsin Card Sorting Test perseverative errors, letter fluency task total correct, animal fluency, and dominant and nondominant mean scores on the Finger Tapping Test). For each score, we converted the raw score to a scaled score, wherein the normative mean was 10 and standard deviation was 3, with higher scores reflecting better performance. We were interested in absolute level of performance, rather than correcting for demographic influences, so these conversions were based on the WAIS-III Reference Group Norms, the 64-card Wis-

TABLE 1. Performance of Older Patients With Schizophrenia on a Neuropsychological Battery (the Aged Schizophrenia Assessement Schedule–Cognitive) at Baseline and at an 8-Week Follow-Up Evaluation[a]

	Baseline		Follow-Up		Analysis			
					Paired sample t test			
Measure	Mean	SD	Mean	SD	t	df	r	Standard Error of Difference
McGurk Visual Spatial Memory Test, total correct (N=42)	53.4	9.6	53.3	9.5	0.07	41	0.59	8.6
Word List Learning Test								
Total learning (N=44)	23.9	5.1	23.5	5.1	0.52	43	0.64	4.3
Delayed recall (N=42)	5.5	1.7	5.3	2.0	0.66	41	0.62	1.5
Trail Making Test[b]								
Part A (seconds) (N=45)	57.4	25.7	57.2	28.4	0.07	44	0.78	17.1
Part B (seconds) (N=43)	178.1	82.2	179.8	88.7	0.17	42	0.72	61.8
Continuous Performance Test								
Total, d-prime (N=37)	1.6	0.7	1.7	1.0	1.09	36	0.63	0.6
Total, omission errors (N=37)[b]	0.6	0.2	0.5	0.3	1.13	36	0.58	0.2
WAIS-III								
Digit span, total raw score (N=45)	12.6	3.1	13.6	4.0	2.24*	44	0.71	2.4
Letter-number sequencing, raw score (N=45)	6.0	2.4	6.3	2.8	0.89	44	0.72	1.8
Digit symbol-coding, total correct (N=45)	37.4	14.8	38.0	14.5	0.57	44	0.89	6.9
Symbol search, raw score (N=45)	16.6	7.8	17.0	9.3	0.50	44	0.75	5.5
Wisconsin Card Sorting Test								
Perseverative errors (N=41)[b]	15.0	8.4	16.0	9.3	0.70	40	0.52	8.3
Categories completed (N=41)	1.4	1.3	1.2	1.4	0.98	40	0.65	1.1
Verbal fluency								
Letter fluency, total correct (N=44)	28.9	10.4	28.6	9.9	0.47	43	0.89	4.9
Animal fluency, total words (N=45)	14.9	5.3	14.9	5.7	0.13	44	0.79	3.4
Finger Tapping Test								
Right hand, total (N=41)	39.7	8.2	40.6	9.3	0.98	40	0.75	5.7
Left hand, total (N=40)	36.8	8.4	37.7	7.6	0.94	39	0.73	6.2
Digit cancellation								
Total correct (N=45)	21.0	9.2	24.3	8.5	3.89***	44	0.81	5.7
Correct minus errors of commission (N=45)	19.4	13.1	23.5	10.4	2.78**	44	0.67	10.7
Benton Judgment of Line Orientation, total correct (N=43)	19.5	6.2	19.4	6.6	0.20	42	0.88	3.0
Hooper Visual Organization Test, total correct (N=45)	20.4	4.7	21.4	4.8	3.76***	44	0.93	1.7
Hiscock Digit Memory Test, total errors (N=44)[b]	0.5	1.3	0.6	1.1	0.10	44	0.32	1.6

[a] Unless otherwise indicated, all t tests were nonsignificant (p>0.05); all test-retest correlations (Pearson's r) were significant. Higher raw scores represent better performance unless otherwise indicated.
[b] Higher raw scores represent worse performance.
*p=0.03. **p=0.008. ***p<0.001.

consin Card Sorting Test Census Matched Adult Sample, and currently in press update of the scaled score conversions from the Heaton, Grant, and Matthews norms. We then calculated the mean scaled score across these tests. We defined those subjects whose mean scaled score was 7 or less as "impaired."

Results

Site Effects

Multivariate analyses of variance (MANOVA) were used to compare the three sites on demographic variables, psychopathology and motor symptoms, and baseline and change scores on the neuropsychological measures. None of these three MANOVAs reached statistical significance.

Demographic Characteristics

The study group represented a diverse range of patients in terms of age (mean=59.4 years, SD=8.4, range=45–77), education (mean=12.0 years, SD=2.6, range=5–17), and ethnicity (Caucasian: 53.3%, N=24; African American: 35.6%, N=16; Latino: 6.7%, N=3; Asian American: 4.4%, N=2). Eighty percent of the patients (N=36) were male. The mean test-retest interval was close to the targeted interval of 8 weeks (actual mean interval was 8.8 weeks, SD=0.9, range=6.9–12.3).

Stability of Psychopathology and Motor Symptoms

There were no significant changes in the mean psychopathology ratings or motor symptoms. All paired-sample t tests comparing the baseline to 8-week follow-up ratings were nonsignificant (all t<1.63, all p>0.11), suggesting no significant changes in mean performance, and all test-retest correlations were significant (all r>0.69, all p<0.001) suggesting participants tended to maintain their relative positions within the distribution of each psychopathology rating or motor rating scale from baseline to follow-up. Since none of the change scores for psychopathology or movement disorder symptoms were significantly different from 0, changes in psychopathology and motor symptoms were not used as covariates in the subsequent analyses evaluating the stability on cognitive measures.

Neuropsychological Performance From Baseline to 8-Week Follow-Up Evaluation

As shown in Table 1, the baseline to 8-week test-retest correlations for each of the 22 neuropsychological scores were all significant (all p<0.001 except for total errors on the Hiscock Digit Memory Test [p<0.04]) suggesting that the performance relative to the entire sample tended to re-

TABLE 2. Rate That Scores on a Neuropsychological Battery (the Aged Schizophrenia Assessement Schedule–Cognitive) at an 8-Week Follow-Up Evaluation Were Worse Than, Within, or Better Than the 90% Confidence Interval of Scores Predicted by Baseline Performance of Older Patients With Schizophrenia and Those Classified as Cognitively Impaired

Measure	Total Number of Subjects	Number of Impaired Subjects	Worse Than CI (%)		Within CI (%)		Better Than CI (%)	
			All Subjects	Impaired Subjects	All Subjects	Impaired Subjects	All Subjects	Impaired Subjects
McGurk Visual Spatial Memory Test, total correct	42	27	2.4	3.7	90.5	88.9	7.1	7.4
Word List Learning Test								
Total learning	44	28	9.1	7.1	88.6	92.9	2.3	0.0
Delayed recall	42	27	7.1	3.4	88.1	93.1	4.8	3.4
Trail Making Test[a]								
Part A	45	29	6.7	3.4	91.1	93.1	2.2	3.4
Part B	43	27	7.0	11.1	88.4	85.2	4.7	3.7
Continuous Performance Test								
d-prime (N=37)	37	22	16.2	0.0	75.7	95.5	8.1	3.5
Omission errors (N=37)	37	22	10.8	4.5	83.8	91.0	5.4	4.5
WAIS-III								
Digit span[a]	45	29	6.7	3.4	86.7	93.1	6.7	3.4
Letter-number sequencing[a]	45	29	2.2	3.4	82.2	82.8	15.6	13.8
Digit symbol-coding[a]	45	29	2.2	3.4	91.1	89.7	6.7	6.9
Symbol search[a]	45	29	6.7	6.9	88.9	89.7	4.4	3.4
Wisconsin Card Sorting Test								
Perseverative errors[a]	41	25	7.3	12.0	85.4	84.0	7.3	4.0
Categories completed	41	25	9.8	8.0	85.4	92.0	4.9	0.0
Verbal fluency								
Letter fluency[a]	44	28	2.3	3.6	93.2	96.4	4.5	0.0
Animal fluency[a]	45	29	8.9	6.9	84.4	86.2	6.7	6.9
Finger Tapping Test[a]								
Right hand	41	26	4.9	7.7	92.7	88.5	2.4	3.8
Left hand	40	25	5.0	8.0	90.0	86.0	5.0	8.0
Digit cancellation								
Total correct	45	29	2.2	3.4	93.3	93.2	4.4	3.4
Correct minus errors of commission	45	29	2.2	3.4	93.3	93.2	4.4	3.4
Benton Judgment of Line Orientation	43	27	9.3	7.4	86.0	88.9	4.7	3.7
Hooper Visual Organization Test	45	29	4.4	55.2	86.7	44.8	8.9	0.0
Hiscock Digit Memory Test	44	29	4.5	0.0	93.2	100.0	2.3	0.0
Summary score[b]	45	29	4.4	6.9	91.1	93.1	4.4	0.0

[a] Task with sufficient normative basis to define impairment and thus included as part of the summary score for defining cognitive impairment in the study.
[b] Performance on the 11 tasks selected to define cognitive impairment in the study. Raw scores were converted to scaled scores with normative mean of 10 (SD=3). A mean scaled score for the 11 tasks of 7 or less was considered impaired.

main stable across the 8-week follow-up. With one exception, all of the correlations were in the range that would be considered a "large effect size" ($r \geq 0.50$). The one exception was the Hiscock score: while statistically significant, the magnitude of the correlation (0.32) was at the lower end of the "medium" effect size range. Yet, constricted variance and a highly skewed distribution of Hiscock errors is to be expected (errors are rare among cooperative examinees), so the potential magnitude of the test-retest correlation may be naturally attenuated. For example, 91% of the sample (N=40 of 44) had 0 or 1 errors and only 4.5% (two patients) had three or more errors, thus truncating the range of scores and reducing the correlations.

As also shown in Table 1, practice effects on most of the 22 neuropsychological test scores were absent or minimal. Paired-sample t tests revealed significant changes on only four of the 22 scores (all in the direction of improved performance with retesting): digit span, digit cancellation (total correct), digit cancellation (total correct minus total errors), and the Hooper Visual Organization Test (total correct). Note that two of these scores came from a single task (digit cancellation), and it is notable that this task was one of the few in the battery that used alternate forms. If the Bonferroni correction were applied to correct for multiple comparisons ($p<0.002$ [0.05/22]), only the digit cancellation total correct and Hooper total correct changes would remain significant.

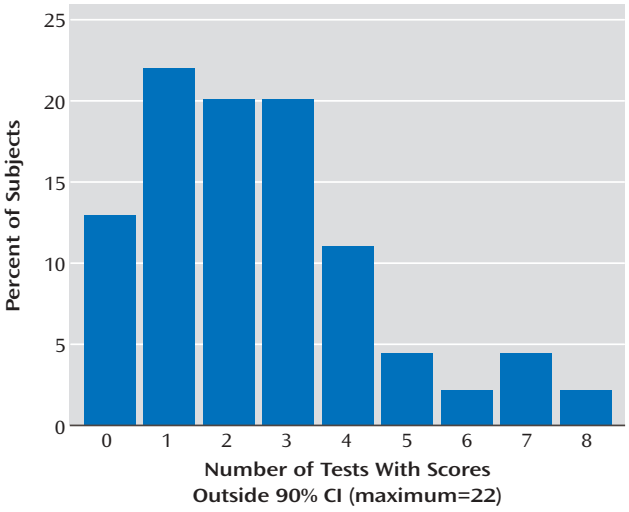
Observed Follow-Up Scores Versus Predicted Follow-Up Scores

For each subject, we calculated the difference between the follow-up score and the baseline score. For each test score, the mean of these differences across all subjects was taken as the expected practice effect for that test. We then added that expected practice effect (the mean difference among all subjects) to each subject's baseline score to compute a "predicted follow-up score." We then placed 90% confidence intervals around each predicted follow-up score using the methods described in Statistical Analyses. For each test, we then computed the proportion of participants whose actual follow-up score was worse than, within, or better than the confidence interval around the predicted follow-up score. As shown in Table 2, most test scores were in fact within the 90% confidence interval

around predicted follow-up scores for each of the 22 test scores. As expected, a small proportion of the observed follow-up scores fell outside the 90% confidence intervals (by chance alone, 5% of the scores are expected to be worse than the 90% confidence interval, 5% are expected to be better than the 90% confidence interval). Since scores of 5% are expected, only those scores substantially greater than 5% reflect a tendency toward excessive change. As seen in the table, with few exceptions, both the impaired patients and the entire sample were essentially similar in the distributions of follow-up scores. The only test with a major skew in its distribution was the Hooper Visual Organization Test, where no impaired patient did notably better on retest and a substantial subset did worse.

Next, for each subject we counted the number of observed follow-up test scores that were worse than, within, or better than the 90% confidence interval around each respective predicted follow-up score. These scores are presented in Figure 1. Only six (13.3%) of the 45 participants had no scores at all with "significant change" (follow-up score better than or worse than the 90% confidence interval) on the 22 tests examined. However, since two tests per subject would be expected to be outside the 90% confidence interval by chance alone, the fact that the majority of subjects (55%) had no more than two follow-up scores that fell outside the 90% confidence interval indicates that these test scores are quite similar in their retest characteristics across subjects. In fact, only 9% of the subjects had more than five individual scores from the 22-test battery that were outside the 90% confidence interval at retest.

FIGURE 1. Number of Tests Within a Neuropsychological Battery (the Aged Schizophrenia Assessement Schedule–Cognitive) in Which Older Patients With Schizophrenia Had Scores at an 8-Week Follow-Up Evaluation That Fell Outside the 90% Confidence Interval of Scores Predicted by Baseline Performance[a]



[a] On average, at least two scores on a 22-score battery would be expected to exceed a 90% confidence interval by random chance.

Discussion

There are several findings of interest in this study. First, there was no evidence of substantial practice-related changes in performance at retest for the majority of the cognitive variables that we examined. This is noteworthy because these are the constructs that are commonly used in clinical trials to study cognitive enhancement with newer antipsychotic medications. Several of these studies have not employed conventional antipsychotic comparator samples because of the difficulty in obtaining patients who are willing to enter a clinical trial and be randomly assigned to older antipsychotic medications. These data suggest that the changes in cognitive functioning reported in those studies, wherein treatment with olanzapine, ziprasidone, and risperidone (28–30) enhanced cognitive functioning, were not likely to be solely due to practice or other retesting effects on these cognitive performance measures when the performance of patients treated with conventional medications is used as a point of reference. In fact, the largest effect size for a retest difference in performance was for the difference between two alternate forms of a single test, the digit cancellation test. This finding suggests, albeit tentatively, that variance across alternate versions of the same test may be greater than retest effects with the same forms in this population.

A second finding of potential importance is the range of variance in correlations between baseline and retest performance. These data may have implications for the ability to detect reliable changes in performance associated with pharmacological or behavioral cognitive enhancement. While most tests had high test-retest reliability, some of the other tests had relatively lower test-retest reliability, such as the Wisconsin Card Sorting Test measures of categories completed and perseverative errors. For tests with reduced test-retest reliability, larger individual changes would be required in order to be detected definitively than the changes seen in some of the more reliable measures. It is of interest that the Wisconsin Card Sorting Test has demonstrated considerable variability in terms of the results across cognitive enhancement studies, while psychomotor speed, which was quite reliable in this study, has been consistently improved with atypical antipsychotic medications across studies.

For some of the variables, however, the high test-retest reliability and low standard error of the retest scores suggests that very small differences in performance would be detectable as clinically significant. For instance, in the domain of visual-motor speed, an improvement in performance of 5 seconds on part A of the Trail Making Test would exceed the 90% confidence interval, and a change of only three items on WAIS-III digit symbol test would exceed this criterion as well. Thus, relatively small changes would be seen as clinically significant, surpassing both practice effects and random variance effects. To put this in

context, in a large-scale clinical trial, patients randomly assigned to treatment with olanzapine improved by an average of 8.5 seconds on part A of the Trail Making Test (28). The findings of the current study would suggest that many of those patients would be expected to have experienced a clinically significant improvement in their trail-making performance as a result.

Recent studies have reported that digit symbol-coding performance from the WAIS-III is the most robust correlate of employment status in patients with schizophrenia (31). The data from this study suggest that this measure has extremely high short-term test-retest reliability. This high reliability also suggests that it would be easier to find correlations between this variable and functional status measures than with other variables (such as the Wisconsin Card Sorting Test) that have relatively somewhat lower test-retest reliability. Similarly, relatively modest changes in performance could be found to be statistically significant because of the high reliability of this measure. At the same time, there are several studies that indicate that digit symbol-coding performance may improve over longer follow-up periods (32, 33). Those studies examined patients much earlier in their illness than the current sample. Conversely, the wide test-retest confidence intervals for some of the other tests indicate that it may be difficult to detect individual changes in performance unless they are quite substantial in magnitude. It should also be considered that the longer the assessment, the more measures that would be expected to improve by chance alone.

A further point that is worthy of note is that the lack of practice effects detected in this study may be associated with the conventional antipsychotic treatments being received by the study participants. Recent research has indicated that treatment with newer antipsychotic medications is associated with greater learning with practice than conventional antipsychotic treatment (34). It has been suggested that conventional antipsychotic treatment may actually interfere with practice effects that would occur otherwise (34, 35). This is a difficult point to prove without retesting patients who are not receiving any antipsychotic treatments. Such a research design is problematic, in that the scientific information to be gained might not be seen to justify the risks to the participants in the study. The small proportion of patients who had any substantial number of test scores outside the 90% confidence interval at the follow-up evaluation suggests that the lack of practice effects was not due to increased random variance in most cases.

The limitations of the study include the age and clinical stability of the patients. Younger patients and those who were acutely ill at entry into the study might have different patterns of stability. Further, tests performed at a ceiling level at baseline could not improve with practice, but this does not appear to be an issue with the instrumentation in this study. There were only two tests in the battery in which baseline performance was even within two standard deviations of perfect scores, the spatial working memory test and the Benton Judgment of Line Orientation test. Thus, 20 of 22 tests still had substantial room for practice effects, and none were actually performed at ceiling level.

In conclusion, the results of this study suggest that older patients with schizophrenia who are receiving conventional antipsychotic treatment manifest quite stable cognitive performance over time, demonstrating little evidence of practice effects and little evidence of wide scatter in retest performance across subjects. There are test-by-test variations in performance that indicate that clinically significant differences across tests may differ. In addition, there may be interindividual differences in retest variance in test scores. This may be an important issue for later research, but the small number of subjects in this study who had any substantial retest variance in their scores would preclude analysis of this factor. Further, the issue of the relative validity of alternative versions of tests versus using the same test across repeated assessments has not been resolved by these results.

Received Dec. 15, 2003; revision received March 9, 2004; accepted March 22, 2004. From the Department of Psychiatry, Mt. Sinai School of Medicine; the Department of Psychiatry, UC San Diego, La Jolla, Calif.; the University of Cincinnati Department of Psychiatry, Cincinnati; and the Department of Psychiatry, University of Indiana School of Medicine, Indianapolis. Address correspondence and reprint requests to Dr. Harvey, Department of Psychiatry, Box 1229, Mt. Sinai School of Medicine, New York, NY 10029; Philipdharvey1@cs.com (e-mail).

This research was funded by an investigator-initiated grant from Eli Lilly and Company, NIMH grant MH-63116 to Dr. Harvey, the VA VISN-III MIRECC, and NIMH grant 5-P30-MH-66248 to Dilip Jeste supporting research at the UC San Diego VA site.

References

1. Harvey PD, Keefe RSE: Studies of cognitive change in patients with schizophrenia following novel antipsychotic treatment. Am J Psychiatry 2001; 158:176–184
2. Keefe RSE, Perkins S, Silva SM, Lieberman JA: The effect of atypical antipsychotic drugs on neurocognitive impairment in schizophrenia: a review and meta-analysis. Schizophr Bull 1999; 25:201–222
3. Friedman JI, Adler DN, Temporini HD, Kemether E, Harvey PD, White L, Parrella M, Davis KL: Guanfacine treatment of cognitive impairment in schizophrenia. Neuropsychopharmacology 2001; 25:402–409
4. Friedman JI, Adler DN, Howanitz E, Harvey PD, Brenner G, Temporini H, White L, Parrella M, Davis KL: A double blind placebo controlled trial of donepezil adjunctive treatment to risperidone for the cognitive impairment of schizophrenia. Biol Psychiatry 2002; 51:349–357
5. Kern RS, Green MF, Marshall BD Jr, Wirshing WC, Wirshing D, McGurk S, Marder SR, Mintz J: Verbal learning in schizophrenia: effects of novel antipsychotic medication. Schizophr Bull 1999; 25:223–232
6. Sax KW, Strakowski SM, Keck PE Jr: Attentional improvement following quetiapine fumarate treatment in schizophrenia. Schizophr Res 1998; 33:151–155

7. Rossi A, Mancini F, Stratta P, Mattei P, Gismondi R, Pozzi F: Risperidone, negative symptoms, and cognitive deficits in schizophrenia: an open study. Acta Psychiatr Scand 1997; 95:40–43
8. Kern RS, Green MF, Marshall BD Jr, Wirshing WC, Marder SR, Mintz J: Risperidone vs haloperidol on reaction time, manual dexterity, and motor learning in treatment-resistant schizophrenia patients. Biol Psychiatry 1998; 44:726–732
9. Lee MA, Thompson PA, Meltzer HY: Effects of clozapine on cognitive function in schizophrenia. J Clin Psychiatry 1994; 55:82–87
10. Temkin NR, Heaton RK, Grant I, Dikmen SS: Detecting significant change in neuropsychological test performance: a comparison of four models. J Int Neuropsychol Soc 1999; 5:357–369
11. Dikmen SS, Heaton RK, Grant I, Temkin NR: Test-retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Test Battery. J Int Neuropsychol Soc 1999; 5:346–356
12. Heaton RK, Temkin N, Dikmen S, Avitable N, Taylor MJ, Marcotte TD, Grant I: Detecting change: a comparison of three neuropsychological methods, using normal and clinical samples. Arch Clin Neuropsychol 2001; 16:75–91
13. Simpson GM, Lindenmayer JP: Extrapyramidal symptoms in patients treated with risperidone. J Clin Psychopharmacol 1997; 17:194–201
14. Beasley CM, Tollefsen G, Tran P, Satterlee W, Sanger T, Hamilton S, Olanzapine (HGAD Study Group): Olanzapine versus placebo and haloperidol: acute phase results of the North American double-blind olanzapine trial. Neuropsychopharmacology 1996; 14:111–123
15. Kay SR: Positive and Negative Syndromes in Schizophrenia. New York, Brunner/Mazel, 1991
16. Montgomery SM: Depressive symptoms in acute schizophrenia. Prog Neuropsychopharmacol 1979; 3:429–433
17. Simpson GM, Lindenmayer JP: Extrapyramidal symptoms in patients treated with risperidone. J Clin Psychopharmacol 1997; 17:194–201
18. Barnes TRE: A rating scale for drug-induced akathisia. Br J Psychiatry 1989; 154:672–676
19. McGurk SR, Meltzer HY: The role of cognition in vocational functioning in schizophrenia. Schizophr Res 2000; 45:175–184
20. Ferris SH, Mackell JA, Mohs R, Schneider LS, Galask D, Whitehouse PJ, Schmitt FA, Sano M, Thomas RG, Ernesto C, Grundman M, Schafer K, Thal LJ: A multicenter evaluation of new treatment efficacy instruments for Alzheimer's disease clinical trials: overview and general results: the Alzheimer's Disease Cooperative Study. Alzheimer Dis Assoc Disord 1997; 11(suppl 2):S1–S12
21. Reitan RM, Wolfson D: The Halstead-Reitan Neuropsychological Test Battery: Theory and Clinical Interpretation, 2nd ed. Tucson, Ariz, Neuropsychology Press, 1993
22. Cornblatt BA, Lenzenweger MF, Erlenmeyer-Kimling L: The Continuous Performance Test, Identical Pairs Version, II: contrasting attentional profiles in schizophrenic and depressed patients. Psychiatry Res 1989; 29:65–85
23. Wechsler D: The Wechsler Adult Intelligence Scale, 3rd ed. San Antonio, Tex, Psychological Corp (Harcourt), 1998
24. Heaton RK, Chelune GJ, Talley JL, Kay GG, Curtiss G: Wisconsin Card Sorting Test Manual. Odessa, Fla, Psychological Assessment Resources, 1993
25. Benton AL, Varney NR, Hamsher KD: Visuospatial judgment: a clinical test. Arch Neurol 1978; 35:364–367
26. Hooper HE: The Hooper Visual Organization Test Manual. Los Angeles, Western Psychological Services, 1983
27. Hiscock M, Hiscock CK: Refining the forced-choice method for the detection of malingering. J Clin Exp Neuropsychol 1989; 11:967–974
28. Harvey PD, Green MF, Meltzer HY, McGurk SR: The cognitive effects of risperidone and olanzapine in patients with schizoaffective disorder or schizophrenia. Psychopharmacology (Berl) 2003; 169:404–411
29. Harvey PD, Napolitano J, Mao L, Gharabawi G: Cognitive enhancing effects of risperidone vs olanzapine in older patients with schizophrenia. Int J Geriatr Psychiatry 2003; 18:820–829
30. Harvey PD, Siu C, Romano S: Randomized, controlled double-blind, multicenter comparison of the cognitive effects of ziprasidone versus olanzapine in acutely ill inpatients with schizophrenia or schizoaffective disorder. Psychopharmacology (Berl) 2004; 172:324–332
31. Gold JM, Goldberg RW, McNary SW, Dixon LB, Lehman AF: Cognitive correlates of job tenure among patients with severe mental illness. Am J Psychiatry 2002; 159:1395–1402
32. Sweeney J, Hass G, Kelp J, Long M: Evaluation of the stability of neuropsychological functioning after acute episodes of schizophrenia: a one year follow-up study. Psychiatry Res 1991; 38: 63–76
33. Mockler D, Riordan S: Memory and intellectual deficits do not decline with age in schizophrenia. Schizophr Res 1997; 26:1–7
34. Harvey PD, Moriarty PJ, Serper MR, Schnur E, Lieber D: Practice-related improvement in information processing with novel antipsychotic treatment. Schizophr Res 2000; 46:139–148
35. Blyler CR, Gold JM: Cognitive effects of typical antipsychotic medication treatment: another look, in Cognition in Schizophrenia. Edited by Sharma T, Harvey PD. Oxford, UK, Oxford University Press, 2000, pp 241–265