Schizophrenia Bulletin vol. 37 no. 2 pp. 334–343, 2011
doi:10.1093/schbul/sbp044
Advance Access publication on June 12, 2009

Validating Measures of Real-World Outcome: The Results of the VALERO Expert Survey and RAND Panel

Feea R. Leifker[2], Thomas L. Patterson[3], Robert K. Heaton[3], and Philip D. Harvey[1,2]

[2]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA 30322; [3]Department of Psychiatry, University of California San Diego Medical Center

Background: **People with schizophrenia demonstrate considerable discrepancy between self-reported functioning and informant reports. It is not clear whether these discrepancies originate from the instruments used or from the perspectives of different informants. The goal of the Validation of Everyday Real-World Outcomes (VALERO) Study is to enhance the measurement of real-world (RW) outcomes in the social, residential, and vocational domains through selection of optimal scales and informants using a multistep process similar to the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) initiative.** ***Methods:*** **Forty-eight experts provided their opinion regarding the best scales measuring RW outcomes. Fifty-nine measures were nominated. The investigators selected the 11 scales that were the most highly nominated, had the most published validity data, and best represented the domains of interest. Information was provided to other experts who served as RAND panelists. Panelists rated each measure for its suitability across multiple a priori domains. Discrepant ratings were discussed until consensus was reached.** ***Results:*** **Following the RAND Panel, the 2 scales that scored highest across the various criteria for each of the classes of scales (hybrid, social functioning, and everyday living skills) were selected for use in the first substudy of VALERO. The scales selected were the Quality-of-Life Scale, Specific Levels of Functioning Scale, Social Behavior Schedule, Social Functioning Scale, Independent Living Skills Schedule, and Life Skills Profile.** ***Discussion:*** **The results show that although there are significant limitations with current scales used for the assessment of RW outcome in schizophrenia, a consensus is possible. Further, several existing instruments were rated as useful for measuring social, residential, and vocational outcomes.**

[1]To whom correspondence should be addressed; tel: 404-727-2707, fax: 404-727-3233, e-mail: philipdharvey1@cs.com

Key words: schizophrenia/cognition/functional outcomes

Background

Deficits in the performance of critical everyday functional skills, including social and occupational functioning, residential maintenance, medication management, and basic self-care, are present in many neuropsychiatric conditions.[1] These impairments are particularly salient in schizophrenia.[2] Disability in schizophrenia occurs even following successful treatment of the clinical symptoms of the illness[3] and often sets in immediately after the first episode.[4] Throughout the course of the illness, the majority of schizophrenic patients experience some form of impairment in everyday functioning, whether in employment, independent living, or social functioning.[5] As a result, disability reduction has the potential to benefit nearly every patient with schizophrenia, yet current treatments for the illness are notably ineffective at reducing disability.[6]

While disability in schizophrenia appears to be related to the failure to perform critical functions in the real world, this disability is likely caused by multiple factors. Failure to perform may be due to skill deficits, motivational deficits, interfering symptoms, and/or limited opportunities or personal resources.[7] Thus, what one does in the real world may not be the perfect index of what one can do, but what one can do under optimal conditions is likely an index of maximal real-world (RW) potential.

We argue that RW functioning is just one element of a more global functional outcomes construct. Factors that influence potential, such as cognitive impairments indexed by neuropsychological (NP) test scores and functional capacity (FC; ie, ability or competence in the performance of everyday living skills), as well as other individual differences such demographic factors and symptoms, including positive, negative, and depressive symptoms, have been shown to predict individuals' RW functioning in schizophrenia.[8] However, reports of RW outcomes vary across informants and contain elements of error, which can be indexed as well. Even ''objective'' milestones such as employment and marriage are influenced by measurable factors other than ability, such

as opportunities and societal incentives and disincentives, and they are often reported inconsistently across informants as described below. Thus, each of the elements of the functional outcomes construct is measured by error-laden indices, and there is no "strict operational" definition of what "RW functional outcome" is.

Arguably, the most consistent element of the functional outcomes construct is NP performance, as measured in recent treatment studies by the MATRICS Consensus Cognitive Battery (MCCB)[8]. The battery was developed through expert nominations from the field and a RAND Appropriateness Panel to select measures in several domains for subsequent comparison in a formal psychometric study.[9–11] The final consensus battery consists of 10 neuropsychological tests and a measure of social cognition, which met comprehensive standards for criterion-referenced validity and test-retest reliability.

Several recent studies have highlighted the variability in convergence between NP, FC, and RW performance measures. In particular, these studies[11–15] examined these convergences in patients with schizophrenia using the University of California San Diego Performance-Based Skills Assessment (UPSA) as the measure of FC. Interestingly, despite the use of different NP performance measures in each of the studies, the correlation between NP performance and the total UPSA score was remarkably consistent, ranging from $r = 0.60$ to $r = 0.65$. Across the same studies, however, the correlation between RW outcomes and UPSA performance varied considerably, ranging from $r = 0.04$ to $r = 0.50$. The studies also showed considerable variance in the correlation between NP performance and RW outcomes: $r = 0.05$ to $r = 0.54$. The lowest correlations were found in studies using only self-report of RW outcomes, and the highest correlation, for both domains, came from a study where the RW outcome used was residential independence measured with a comprehensive assessment involving multiple sources of information. These data suggest 2 conclusions: First, performance-based measures of NP performance and FC are highly convergent with each other regardless of the NP battery employed; and second, different RW outcome measures yield widely variant correlations with corresponding performance-based measures. Because the correlations between the performance-based NP and FC measures were so consistent, the variation in correlations with RW measures and these other domains implicates shortcomings of the RW outcomes measures.

The overriding goal in treating cognitive deficits in schizophrenia is reducing functional disability. However, if the overlap between NP performance, even if measured with a highly reliable and valid assessment battery, and RW outcome is as small as it initially appears, the question remains as to whether successful treatment of cognition can realistically improve RW outcomes. One suggestion, explored below, is that current instruments assessing RW outcomes exhibit intrinsic limitations, at least when in the hands of certain informants. The most global and arguably most significant aspects of RW outcome can be measured with high reliability, admittedly with certain limitations. These include marriage or an equally stable relationship, full-time employment, and self-supported living. However, these outcomes are rare and develop over time; hence, they are impractical for use as outcomes variables in treatment studies, even for trials of treatment effectiveness. Measurement of more subtle aspects of RW outcome in neuropsychiatric conditions (eg, household management, social contacts, and job seeking activities) is rarely direct, and in many research studies, these aspects are often measured through self-report. However, recent research has shown that self-reports by patients with schizophrenia may be unreliable when compared with other sources of information; schizophrenia seems to induce types and degrees of self-report deficits that exceed those of the general population.[16] Patients with schizophrenia manifested substantial problems in self-reporting their cognitive impairments, when examined on a structured rating scale that was then related to their performance on an NP assessment.[17] Further, the convergence of case manager reports and patient self-reports, even of supposedly objective outcomes such as living situation and time spent working in the past week has been found to be minimal, accounting for as little as 4% of joint variance.[18] Patients' self-report of their functioning in that same study was not as strongly associated with their performance on NP and FC measures than were case manager reports, suggesting that these case manager reports have evidence of more validity than patients self-report.

The modifiable sources of reduced validity for rating RW outcomes are at least 2-fold: first, the characteristics of the informant used and, second, the RW outcome rating scale selected. Variance in reports by informants can be influenced by the amount of contact with the subject and situation specificity of the observation. In the case of self-report, the variation can be influenced by patients' competence in self-evaluation of the quality as well as the quantity of their performance (see Bowie et al[18] for an example of this). It is entirely possible that a substantially greater correlation exists between NP performance and aspects of RW outcomes than has been detected in previous studies where the RW outcome measures may have been deficient. For example, in the Twamley et al[15] study, the RW outcome was based on a comprehensive assessment of residential independence, and the correlation between NP performance and this outcome was the highest for any of the studies cited above. Therefore, the next step in the construct validation process would be to evaluate candidate measures of RW outcomes with rigorous process similar to the selection of the MCCB.

Overview of Validation of Everyday Real-World Outcomes

The Validation of Everyday Real-World Outcomes (VALERO) in schizophrenia project represents a joint effort between researchers at Emory University and the University of California, San Diego. The main goal of the project is to improve the assessment of RW functioning and hopefully apply the findings to future treatment studies of schizophrenia. To do so, VALERO will examine the convergence between a wide range of existing RW rating scales with performance-based measures, including NP test scores and FC assessments. Researchers will identify the existing RW outcomes scale (or subscales from existing scales) that is most highly convergent in a longitudinal design. Next, the identity of the informant whose ratings are most convergent with the rest of the outcomes construct will be investigated. Candidate informants include patient self-report, a relative/caregiver, a case manager or other high-contact clinician, and a medical prescriber. Further, the VALERO project will systematically study factors possibly associated with discrepancies between self-appraisal and informant appraisal of RW functional outcomes (such as depression, metacognitive skills, and emotional intelligence) in order to inform later research attempting to increase congruence of appraisals.

The VALERO Study will complete these goals in 3 substudies. Study 1 will use assessment scales selected by a RAND Appropriateness Panel to obtain RW functional status ratings and examine the convergence of those ratings with each other and with NP and FC scores. Study 2 will attempt to determine the best informant of patient functioning, and Study 3 will examine how demographic factors, psychiatric symptoms, and other features of illness affect the convergence of self-report and informant ratings of patients' RW functional skills performance.

In this article, we report the first step in this process. The current study used an expert survey and RAND Panel, such as those employed in the MATRICS process,[9] to select the most suitable current RW outcomes measures for entry into the validation study.

Methods

Expert Panel Nominations

A list of experts was compiled by the grant authors based on personal experience, literature searches and networking connections. Further, feedback on the expert list was provided by two rounds of review by a National Institute of Mental Health (NIMH) study section. The experts were selected because they conducted research or performed high-level clinical activity in an area that would inform the nomination process and for the breadth of types of activities in which they were engaged. Researchers and leading clinicians in academia, the pharmaceutical industry, and in rehabilitation medicine and occupational therapy were surveyed, as the ultimate goal of this project is to inform outcome measurements in a large-scale cognitive enhancement trial.

In September of 2007, e-mails sending overviews of the study and defining the concept of everyday outcomes as operationalized in this study were sent to 46 researchers and professionals. These experts were asked to "nominate the scales that you think best measure everyday outcomes in schizophrenia. The outcomes may include social, vocational, independent living, self-care, or any combination of these." In addition, the 9 individuals selected to compose the RAND Panel were also asked to submit their own nominations. The nomination process concluded in November 2007 after each expert received 2 reminder e-mails.

Upon conclusion of the nomination process, the investigators (Drs P.D.H., R.K.H., and T.L.P.) examined the most frequently nominated scales and identified all those that met the a priori criteria for continuation to the next stage. These broad criteria were that (*a*) the scale was nominated by the experts surveyed, (*b*) the scale had available data (published or unpublished) regarding its psychometric qualities, and (*c*) the scale assessed social functioning, everyday living skills, or both these areas ("hybrid" scales).

Selection Criteria for RAND Panel

Once the investigators eliminated ineligible scales for review at the RAND Panel, they established the various scale characteristics, which would be provided to and rated by the panelists. The characteristics chosen were similar to those deemed important in the MATRICS process.[19] The entire citation history of the original published article for each scale was retrieved from 2 search engines: Web of Knowledge and Google Scholar. All articles citing the scale were retrieved and examined for information regarding the domains chosen by the investigators. The final rating domains selected were reliability (test-retest and interrater), convergence with other measures of the functional outcomes construct: performance-based measures of FC, and NP performance, sensitivity to treatment effects, usefulness for multiple informants (eg, self, friend or relative, case manager, or prescriber), relationships with symptom measures, practicality and tolerability for people with low education levels, and convergence with other measures of RW functional outcomes (including either other rating scales or achievement milestones). Final definitions of each of these domains are described in the "Results" section.

Data Preparation and Transmission and RAND Process

Data in these areas were compiled in a summary sheet along with a brief description of the scale that included

time it took to administer, reference period of RW functioning, and additional information describing how the scale should be administered. These data along with copies of each scale and the citation history were distributed to the 8 panelists and chairperson (Stephen R. Marder, MD). The representatives of the panel included schizophrenia researchers studying functional outcomes, providing psychosocial treatments for disability, and conducting pharmacological treatment studies and experts on pharmacological treatments (see Appendix for a full list). No member of the panel reported a real or potential conflict of interest with the outcome of the process.

All panelists were given 1 month to review the information and were asked to submit preliminary ratings on the scales before they met at the Panel. Scales were rated on a 9-point (1–9) scale, where scores of 1–3 were poor, 4–6 were fair to good, and 7–9 were very good to superb. Preliminary results were compiled in each scale domain for each of the functional outcomes scales. These results were assembled into summary tables for each scale showing the mean, range, and SD of the preliminary survey results. These summary tables were provided to the panelists at the RAND Panel meeting.

The RAND Panel meeting was open to NIMH staff and other interested parties, with only scale developers being recused. During the RAND Panel meeting, 2 NIMH observers attended the panel but did not submit formal ratings of the scales. The panel focused on resolving discrepant ratings. Panelists discussed each item for each rating scale that had an SD of greater than 2 points until a consensus of a 1-point range around a mean value could be reached (ie, rating of 3 ± 1). Panelists then submitted their final ratings within this range.

Results

Expert Panel Nominations

Thirty-one e-mails led to a usable nomination (67.4% response rate) from the expert nominators, and an additional 7 experts returned an e-mail declining to participate or referring us to contact someone else. Five of the total 38 people who returned e-mails containing nominations were in the pharmaceutical industry, while the remaining 33 were in academia. Of those nominations returned, 27 e-mails contained nominations of scales that met the general criteria of the investigators.

Upon conclusion of the nominations, the experts surveyed had suggested 59 different measures. The investigators selected 2 hybrid measures, 2 social functioning measures, and 5 everyday living scales that they felt best met the agreed upon criteria and for which the literature search process was conducted. These scales are described briefly in Table 1, and their primary citation and nomination history are listed.

RAND Panel

Following the panelists' preliminary review of the scale information, descriptive statistics regarding their opinions was compiled in each domain for each of the 9 rating scales. These data are shown in table 2. In the initial ratings, panelists had disagreement, noted by an SD $>$ 2, on 6 items. It was noted at the meeting that discrepancy was more prevalent in the domains of practicality, usefulness for multiple raters, and comprehensiveness. These domains varied significantly due to incongruence in the panelists' personal perception of the definitions in each of these domains. Significant time was spent during the meeting refining definitions of these areas. Therefore, each of these domains, regardless of degree of discrepancy, was examined during the review of the scales to ensure that the ratings matched the revised definitions (see table 3). Following a clarification of the definitions, the panelists reached consensus on all items. No mean scale rating at the close of the panel differed significantly from the original rating (at $P < .05$).

During the panel, it was determined that one of the everyday living scales, the Role Functioning Scale (RFS), acted as a summary scale rather than an actual rating instrument. This determination occurred because analysis of the scale showed that it just assessed global functioning in 4 broad areas of functional outcomes rather than asking specific questions regarding the patients' functioning in specific areas. As a result, it was decided that the scale would be excluded from the panel conversation. The final descriptive statistics of the panel's consensus ratings are presented in table 4.

Description of Scales Selected

Quality-of-Life Scale. The Quality-of-Life Scale (QLS) is a 21-item semistructured interview assessing functioning in schizophrenia. The scale addresses functioning across 4 domains: (1) intrapsychic foundations, (2) interpersonal relations, (3) instrumental role category, and (4) common objects and activities. The QLS is administered by a trained interviewer or clinician and takes about 45 minutes to complete. The scale assesses functioning over the past 4 weeks. Each of the 21 items is rated based on the interviewers' opinions of the patient's functioning. The interviewer rates the patient on a 7-point scale with higher scores indicating higher levels of functioning. Scores on each of the items in a domain can be summed to create a subscale score, and all items can be summed to create an overall score on the QLS that ranges from 0–126.

Specific Levels of Functioning Scale. The Specific Levels of Functioning (SLOF) Scale is a 43-item multidimensional behavioral survey administered in person to the

Table 1. Scales Selected to Be Reviewed by the RAND Panel of Experts

Scale	Abbreviation	VALERO Classification	Original Citation	Total Citations[a]	Total Useable Articles	Number of Expert Nominations
Heinrichs-Carpenter Quality-of-Life Scale	QLS	Hybrid	Heinrichs et al[20]	512	30	8
Specific Levels of Functioning Scale	SLOF	Hybrid	Schneider and Strening[21]	34	7	3
Multidimensional Scale of Independent Functioning	MSIF	Hybrid	Jaeger et al[22]	13	4	2
Birchwood Social Functioning Scale	SFS	Social functioning	Birchwood et al[23]	134	17	9
Social Adjustment Scale II	SAS-II	Social functioning	Schooler et al[24]	32	7	3
Social Behavior Schedule	SBS	Social functioning	Wykes and Stuart[25]	179	17	3
Multnomah Community Ability Survey	MCAS	Everyday living	Barker et al[26,27] and Dickerson et al[28]	75	7	11
Life Skills Profile	LSP	Everyday living	Rosen et al[29]	136	7	5
Independent Living Skills Survey	ILSS	Everyday living	Wallace et al[30]	25	7	4
Independent Living Skills Inventory	ILSI	Everyday living	Menditto et al[31]	13	3	3
Role Functioning Scale	RFS	Everyday living	Goodman et al[32]	45	13	4

Note: VALERO, Validation of Everyday Real-World Outcomes.
[a]At the time of the scale literature search (Fall 2007).

caseworker or caregiver of a schizophrenic patient. The scale assesses the patient's current functioning and behavior across the following 6 domains: (1) physical functioning, (2) personal care skills, (3) interpersonal relationships, (4) social acceptability, (5) activities of community living, and (6) work skills. Each of the questions in the above domains is rated on a 5-point Likert scale. Scores on the instrument range from 43 to 215. The higher the total score, the better the overall functioning of the patient. The exact time frame that the survey attempts to assess functioning for is unspecified. The scale also includes an open-ended question asking the informant if there are any other areas of functioning not covered by the instrument that may be important in assessing functioning in this patient. Each informant is asked to rank how well they know the patient on a 5-point Likert scale ranging from "not well at all" to "very well."

Social Behavior Schedule. The Social Behavior Schedule (SBS) is a 30-item measure used to assess social functioning in chronic (community or hospital) psychiatric populations. The survey assesses a patient's past month functioning in 21 areas. The scale is administered as a semistructured interview and is given to an informant. The scale takes approximately 15 minutes to deliver. Most items are rated on a 5-point scale, with a higher score representing lower levels of functioning. Scores in each of the 21 areas can be used alone as indicators of functioning, or a total SBS score can be used. In addition, 2 additional scores can be derived from the scale, the severe behavior problems score (BSS) and the mild and severe behavior problems score (BSM). Behaviors rated areas 3 or 4 in the 21 areas are rated BSS, and items that are rated as 2, 3, or 4 are rated as BSM.

Social Functioning Scale. The Birchwood Social Functioning Scale (SFS) was developed to assess social adjustment in schizophrenic patients. The 79-item measure assesses social functioning across 7 domains: (1) social engagement/withdrawal, (2) interpersonal behavior/communication, (3) prosocial activities, (4) recreation, (5) independence—competence, (6) independence—performance, and (7) employment/occupation.

The SFS takes approximately 30–45 minutes to administer and can be used as a self-report or informant interview, although it is generally administered to an informant. Items are scored on a 4-point scale

Table 2. Statistics for Panelists' Preliminary Rating of Functional Outcomes Scales

	Domains Rated, M (SD)							
	Reliability	Convergence	Sensitivity	Practicality	Usefulness for Multiple Raters	Relationship With Symptoms	Comprehensiveness	Scale Mean Score
QLS	7.63 (0.92)	4.38 (1.41)	6.88 (1.46)	5.50 (1.93)	4.63 (2.07)	5.25 (1.91)	6.25 (1.49)	5.79 (0.87)
SLOF	3.25 (1.04)	5.63 (1.30)	6.75 (0.89)	6.63 (1.77)	4.63 (1.51)	4.88 (0.83)	5.75 (1.91)	5.36 (0.74)
MSIF	6.88 (1.36)	5.38 (1.19)	4.38 (0.52)	5.38 (2.00)	4.13 (1.89)	5.13 (1.13)	6.00 (0.93)	5.32 (0.57)
SFS	5.88 (1.13)	3.75 (0.89)	6.00 (1.31)	5.25 (1.67)	6.25 (1.04)	5.00 (1.20)	6.63 (1.19)	5.54 (0.60)
SAS-II	4.88 (1.55)	3.50 (1.31)	6.50 (0.76)	4.38 (1.69)	3.50 (1.60)	4.63 (1.19)	6.00 (1.51)	4.77 (1.07)
SBS	6.63 (1.19)	5.25 (1.04)	6.50 (1.31)	6.38 (1.51)	7.00 (1.20)	4.13 (1.73)	5.25 (2.49)	5.88 (0.90)
MCAS	5.88 (1.55)	4.50 (0.93)	5.25 (1.28)	5.00 (1.93)	4.75 (1.39)	3.75 (1.28)	5.00 (2.14)	4.88 (0.77)
LSP	5.50 (0.93)	3.50 (0.76)	5.50 (1.41)	5.75 (1.16)	5.88 (1.36)	4.0 (0.93)	4.63 (1.41)	4.96 (0.71)
ILSS	4.75 (0.89)	3.25 (1.28)	3.75 (1.39)	5.25 (1.67)	5.13 (1.13)	5.38 (2.26)	6.63 (1.19)	4.88 (0.60)
ILSI	4.75 (0.89)	5.88 (0.83)	4.13 (0.35)	4.5 (1.77)	3.50 (1.93)	4.88 (1.13)	5.75 (1.83)	4.77 (0.46)
RFS	5.13 (1.25)	2.88 (1.46)	6.50 (1.20)	5.38 (2.20)	3.38 (1.41)	4.63 (2.26)	3.00 (1.31)	4.41 (0.59)
Domain mean score	5.56 (1.15)	4.35 (1.13)	5.65 (1.08)	5.40 (1.75)	4.80 (1.50)	4.69 (1.44)	5.53 (1.58)	

Note: QLS, Quality-of-Life Scale; SLOF, Specific Levels of Functioning; MSIF, Multidimensional Scale of Independent Functioning; SFS, Social Functioning Scale; SAS-II, Social Adjustment Scale II; SBS, Social Behavior Schedule; MCAS, Multnomah Community Ability Survey; LSP, Life Skills Profile; ILSS, Independent Living Skills Survey; ILSI, Independent Living Skills Inventory; RFS, Role Functioning Scale.

with higher scores indicating a higher level of functioning. Raw scores on each of the subscales are converted to a scale score. The reference period for this scale is unspecified.

Life Skills Profile. The original version of the Life Skills Profile (LSP) is a 39-item informant survey assessing a patient's level of functioning. Family members, psychiatric professionals, or case workers can be used as the informant in the interview. Multiple informants can be used to create a mean informant score for each patient. The scale assesses functioning in 5 areas: (1) self-care, (2) nonturbulence, (3) social contact, (4) communication, and (5) responsibility.

Items are rated on a 4-point scale with higher scores reflecting lower functioning. The mean of the scores in each subscale is used to represent a patient's functioning in each of this areas. The time frame in which the LSP is used to assess functioning is unspecified; however, most studies have used a 3-month range.

Independent Living Skills Survey. The Independent Living Skills Survey (ILSS) is a checklist measure of basic functioning for individuals with severe and persistent mental illnesses. There are 2 versions of the ILSS, the self-report version (ILSS-SR) and the informant version (ILSS-I). Both versions can be administered in person or on paper and rate the patients on their functioning over the past 30 days.

The ILSS-I is a 103-item scale assessing basic community living skills such as appearance and care of clothing, personal hygiene, care of personal possessions and living space, food preparation, eating behaviors, care of one's own health and safety, money management, transportation, leisure and recreational activities, job seeking, job maintenance, and social interactions. Informants rate the patient on a 5-point scale ranging from never to always. The ILSS-I takes from 20 to 35 minutes to administer. The average score of each functional area is computed to determine the overall level of functioning in a given area where higher scores will mean higher functioning.

The ILSS-SR is a 61-item scale measuring appearance and care of clothing, personal hygiene, care of personal possessions, food preparation and storage, health maintenance, money management, transportation, leisure and community, job seeking, and job management. If given in interview format, there are 9 questions for the interviewer to respond to regarding the appearance of the patient. Similar to the ILSS-I, patients are asked to rate whether or not (yes or no) they complete basic tasks. Answers are summed (no = 0, yes = 1) and averaged per area. The ILSS-SR takes approximately 20–30 minutes to administer.

Table 3. Final Domains Rated by VALERO RAND Panelists and Their Definitions and Suggestions for Aspects to Consider in Rating Each Domain

Domain	Definition/Explanation of Term
Reliability	Assessment of test-retest reliability (does the measure produce the same distribution with repeated ratings?) and interrater reliability (across "similar" raters such as clinicians)
Convergence with NP and FC performance	Do scores on these rating scales converge well with objective data from performance-based assessments of both cognitive functioning and social and everyday living skills? Are correlations consistent across studies (and across raters, if available)? Scales where there are positive data available for both FC and NP performance should receive higher ratings. Convergence with multiple cognitive domains is more desirable than single ones; the MATRICS NP battery is fully representative of the important domains for the current study.
Sensitivity to change	Sensitivity to change examines whether the scale is constructed in such a way that it would be possible to detect changes in real-world functional status. Does the scale define functioning in a way consistent with detection of changes? Scales that do not separate lifetime and current functioning or scales that measure personality traits would be rated more poorly.
Practicality and tolerability	This domain is concerned with whether the rating scale is accessible to all potential informants. Does completing the rating scale require a glossary (or a dictionary)? Length of the scale and clarity of the item definitions should strongly affect these ratings. Any rating scale that could be reliably completed without a formal interview process would receive considerably higher ratings. Also, scales that could reliably be conducted as self-report measure or completed in a short duration of time would score higher on this domain.
Usefulness for multiple informants	The domain assesses whether the scale, as currently configured, would be directly useful for multiple raters who know the patient in different ways, or would only certain informants know the information required? These raters could include the patient, a high-contact clinician, close friend or relative, or a medical practitioner. Scales with predefined alternate forms for self-report and informant report would get higher ratings as would scales with empirical evidence that the scale can be used with validity by multiple raters.
Relationship with symptom measures	Relationship with symptom measures should be assessed through data regarding the correlations between scores on these functional measures and symptom severity. Are scores on these measures excessively influenced by symptoms? Scales rated more highly may be scales where there is evidence of moderate correlation with symptom measures but no indications that scores on the scale are a simply proxy for the severity of other symptoms, such as psychosis. Regarding negative Symptoms, correlations should be low to moderate so that scales are not overlapping with symptom measures but do reflect the relationship between negative symptoms and functional outcomes.
Comprehensiveness of assessment	This domain attempts to determine the extent to which the scale provides a comprehensive measure of what it attempts to examine. How well does the scale assess the domains of interest? Some rating scales measure only social or everyday living outcomes. Ratings should not be reduced on that basis but rather adjusted to consider the domain involved. A hybrid scale that is good at rating some elements of outcome (everyday functioning) and weak on social functioning would receive lower ratings.

Note: VALERO, Validation of Everyday Real-World Outcomes; NP, neuropsychological; FC, functional capacity; MATRICS, Measurement and Treatment Research to Improve Cognition in Schizophrenia.

Discussion and Implications

Preliminary Ratings

An examination of the preliminary ratings of the scales shows discrepancies on certain measures and domains. In general, there was greater variance on the ratings of the measures in the domains of practicality, comprehensiveness, and usefulness for multiple raters, suggesting variability in interpretations of the definitions as discussed above as well as variation across the scales in their approaches to assessments of RW outcomes. The social functioning domain (which included the SFS, SBS, and Social Adjustment Scale II) showed the most amount of variability in the mean score of each measure's ratings. This could suggest that the panelists had the hardest time fitting these measures into the domains or that the panelists differentially rated the utility of the various social functioning measures. The mean scores on the social scales, however, are quite high, suggesting that the panelists felt that they were as useful as the hybrid scales. The mean measure rating on the everyday living scales were often rated at least a point lower (the lowest being the RFS) when compared with the highest rated social and hybrid scales (SBS and QLS, respectively). Unlike the SFSs, there was little variability in the SDs of the mean scale ratings in this domain, suggesting that panelists

Table 4. Final Domain Ratings by RAND Panelists

	Domains Rated, M (SD)							
	Reliability	Convergence	Sensitivity	Practicality	Usefulness for Multiple Raters	Relationship With Symptoms	Comprehensiveness	ean Scale Score
QLS	7.63 (0.92)	4.25 (1.16)	6.05 (1.77)	5.25 (0.89)	4.88 (0.83)	4.75 (0.71)	6.38 (1.19)	5.66 (0.44)
SLOF	3.25 (1.04)	5.63 (1.30)	6.75 (0.89)	6.13 (1.25)	4.50 (1.31)	4.88 (0.83)	4.63 (0.92)	5.11 (0.68)
MSIF	6.88 (1.36)	5.38 (1.19)	4.38 (0.52)	4.75 (1.04)	4.75 (0.71)	5.13 (1.13)	6.00 (0.93)	5.32 (0.39)
SFS	5.88 (1.13)	3.75 (0.89)	6.00 (1.31)	5.63 (1.06)	6.25 (1.04)	5.00 (1.20)	6.63 (1.19)	5.59 (0.55)
SAS-II	5.25 (1.04)	3.50 (1.31)	6.50 (0.76)	4.50 (0.53)	3.38 (0.74)	4.63 (1.19)	5.88 (0.83)	4.80 (0.66)
SBS	6.63 (1.19)	5.25 (1.04)	6.50 (1.31)	5.88 (1.25)	7.00 (1.20)	3.13 (0.64)	4.25 (1.75)	5.52 (0.60)
MCAS	5.88 (1.25)	4.50 (0.93)	5.25 (1.28)	4.38 (1.19)	6.63 (1.06)	3.75 (1.28)	4.25 (1.04)	4.66 (0.67)
LSP	5.50 (0.93)	3.50 (0.76)	5.50 (1.41)	5.75 (1.16)	5.88 (1.36)	4.00 (0.93)	4.63 (1.41)	4.96 (0.71)
ILSS	4.75 (0.89)	3.25 (1.28)	3.75 (1.39)	5.25 (1.49)	5.13 (1.13)	5.25 (1.28)	6.63 (1.19)	4.86 (0.47)
ILSI	4.75 (0.89)	5.88 (0.83)	4.13 (0.35)	4.00 (0.76)	2.13 (0.64)	4.88 (1.13)	6.13 (0.99)	4.55 (0.14)
Mean Domain Score	5.64 (1.06)	4.49 (1.07)	5.53 (1.10)	5.15 (1.06)	4.85(1.00)	4.54 (1.03)	5.54 (1.14)	

Note: Scaling for ratings is as follows: 1, poor; 3, fair; 5, good; 7, very good; 9, superb. QLS = Quality-of-Life Scale; SLOF = Specific Levels of Functioning Scale; MSIF = Multidimensional Scale of Independent Functioning; SFS = Social Functioning Scale; SAS-II = Social Adjustment Scale II; SBS, Social Behavior Schedule; MCAS = Multnomah Community Ability Survey; LSP = Life Skills Profile; ILSS = Independent Living Skills Scale; ILSI = Independent Living Skills Inventory.

may have possessed similar overall opinions of these scales.

Final Ratings

The final panel ratings followed a trend similar to the preliminary panel ratings. The mean scale ratings of hybrid and SFSs were often higher than the everyday living scales, although less so than during the preliminary ratings. Ratings of the LSP remained exactly the same from pre to post panel ratings. The ratings of the SFS and ILSS only changed in one domain (practicality and symptoms, respectively) and only slightly so. Interestingly, the ILSS and LSP were the highest rated everyday living scales at the preliminary ranking and stayed so upon final rating. Overall, the QLS scored most highly in the final ratings over all constructs. The SFS and LSP scored the highest in their respective constructs. No particular domain appeared to have more variance than others during the final rating.

Scale Selection

Following review of the final ratings, the investigators selected 2 scales from each of the classifications (hybrid, everyday living, and social functioning) to be used in the first validation study. The Heinrichs-Carpenter QLS and SLOF Scale were selected for validation. Although the Multidimensional Scale of Independent Functioning (MSIF) exhibited higher ratings than the SLOF at the conclusion of the RAND Panel, the investigators opted to use the SLOF because the RAND Panel noted that the MSIF had been used reliably with bipolar patients but lacked extensive data on patients with schizophrenia and that their ratings were based on these bipolar data. The 2 social functioning measures with the highest ratings by the experts, the Birchwood SFS and the SBS, were selected to represent that domain. Likewise, the LSP and ILSS were the highest rated in their construct and were chosen to represent the everyday living skills scales. Interestingly, the Multnomah Community Ability Scale was the most frequently nominated scale by the experts but was not the most highly rated by our panelists, some of whom had used this scale in their research. Failing to rate on popularity may show a lack of bias on the part of the panelists.

The results of this study reflect the current consensus in the field with regard to functional outcomes scales. No scale received a mean total score rating over 6 or below 4—suggesting that all current scales are viewed as moderately useful in their current versions, with some meeting minimal criteria for acceptability for use as currently configured. The origin of these ratings was not based on poor performance in previous studies. Rather, many of these scales lack critical data regarding basic reliability across raters and relationships with other elements of the functional outcomes construct, including NP and FC performance. Further, although each of the selected scales has evidence of sensitivity to RW

milestones, such as independent living and social outcomes, many of the previous studies used very broad indices of these outcomes (institutionalized vs ambulatory) as the outcomes variables. Ratings for usefulness across multiple raters were also quite low, partly because many of these scales do not have alternate forms that attempt to capture the differing perspectives of different raters. The panelists' consensus then indicates that those in the field of schizophrenia research have not yet determined an entirely effective measure of the RW outcomes component of the functional outcomes construct but that some measures are likely to be suitable in the interim. The VALERO Study will attempt to determine the best scale or compilation of subscales in order to create an RW functional outcomes measure that could serve as an outcome measure in future clinical trials concerned with improving cognition and functional disability in patients with schizophrenia. This first component of the VALERO project demonstrates the need for a scale that can score highly on all the domains investigated in this study and also serve as a practical and informative outcome measure in the area of schizophrenia research. The first phase of the VALERO Study will in fact directly examine the 2 most problematic aspects of this current group of scales: their temporal stability and reliability as well as the usefulness across multiple informants who report on the same person with schizophrenia.

Funding

National Institute of Mental Health (linked grants NIMH MH78775 to P.D.H., MH78737 to T.L.P.).

Appendix A

RAND Panel Committee Members

- Steven R. Marder (University of California, Los Angeles), chairperson
- Alan Bellack (University of Maryland, Baltimore Veterans Affairs)
- George Garibaldi (Hoffmann La Roche)
- Terry Goldberg (Zucker Hillside Hospital)
- Eric Granholm (University of California, San Diego, Veterans Affairs)
- Robert Kern (University of California, Los Angeles)
- Antony Loebel (Dainippon Sumitomo Pharma America)
- Anthony Stringer (Emory University)
- Dawn Velligan (University of Texas Health Science Center at San Antonio)

Acknowledgments

In the past 3 years, Dr Harvey has served as an advisor or consultant to: Astra-Zeneca Pharmaceuticals; Dainippon Sumitomo Pharmaceuticals America; Eli Lilly and Company; Johnson and Johnson, Inc; Merck and Company; Novartis Pharmaceuticals; Pfizer, Inc; SolvayWyeth Alliance; and the Sanofi-aventis group. Dr Harvey received grant or contract support from Astra-Zeneca Pharmaceuticals and Johnson and Johnson, Inc. All other authors report no other outside relationships. Articles in the reference section numbered from 20 to 32 describe nominated scales.

References

1. Murray CJL, Lopez AD. Global mortality, disability, and the contributions of risk factors: global burden of disease study. *Lancet*. 1997;349:1436–1442.
2. Wiersma D, Wanderling J, Dragomirecka E. Social disability in schizophrenia: its development and prediction over 15 years in incidence cohorts in six European centres. *Psychol Med*. 2000;30:1155–1167.
3. Robinson DG, Woerner MG, McMeniman M, Mendelowitz A, Biler RM. Symptomatic and functional recovery from a first episode of schizophrenia or schizoaffective disorder. *Am J Psychiatry*. 2004;161:473–479.
4. Ho BC, Andreasen N, Flaum M. Dependence on pubic financial support early in the course of schizophrenia. *Psychiatr Serv*. 1997;48:948–950.
5. Harvey PD, Green MF, Keefe RSE, Velligan D. Cognitive function in schizophrenia: its role in the definition and evaluation of effective treatments for the illness. *J Clin Psychiat*. 2004;65:361–372.
6. Hegarty JD, Baldessarini RJ, Tohen M, Waterneaux C, Oepen G. One hundred years of schizophrenia: a meta-analysis of the outcome literature. *Am J Psychiatry*. 1994; 151:1409–1416.
7. Harvey PD, Velligan DI, Bellack AS. Performance-based measures of functional skills: usefulness in clinical treatment studies. *Schizophr Bull*. 2007;33:1138–1148.
8. Bowie CR, Leung WW, Reichenberg A, et al. Predicting schizophrenia patients' real world behavior with specific neuropsychological and functional capacity measure. *Biol Psychiatry*. 2008;63:505–511.
9. Neuchterlein KH, Green MF, Kern RS, et al. The MATRICS Consensus Cognitive Battery, part 1: test selection, reliability, and validity. *Am J Psychiatry*. 2008;165:203–213.
10. Kern RS, Nuechterlein KH, Green MF, Baade LE, Fenton WS, et al. The MATRICS Consensus Cognitive Battery, part 2: co-norming and standardization. *Am J Psychiatry*. 2008;165:214–220.
11. Green MF, Nuechterlein KH, Kern RS, et al. Functional co-primary measures for clinical trials in schizophrenia: results from the MATRICS Psychometric and Standardization Study. *Am J Psychiatry*. 2008;165:221–228.
12. McKibbin C, Patterson TL, Jeste DV. Assessing disability in older patients with schizophrenia: results from the WHO-DAS-II. *J Nerv Ment Dis*. 2004;192:405–413.
13. Bowie DR, Reichenberg A, Patterson TL, Heaton RK, Harvey PD. Determinants of real world functional performance in schizophrenia: correlations with cognition, functional capacity, and symptoms. *Am J Psychiatry*. 2006;163:418–425.
14. Keefe RSE, Poe M, Walker T, Harvey PD. The relationship of the Brief Assessment of Cognition in Schizophrenia (BACS) to functional capacity and real-world functional outcome. *J Clin Exp Neuropsychol*. 2006;28:260–269.

15. Twamley EW, Doshi RR, Nayak GV, Palmer B. Generalized cognitive impairments, ability to perform everyday tasks, and level of independence in community living situations of older patients with psychosis. *Am J Psychiatry*. 2002;159: 2013–2020.
16. World Health Organization. International Pilot Study of Schizophrenia. Author: Geneva, Switzerland, 1973.
17. Keefe RSE, Poe M, Walker TM, Kang JW, Harvey PD. The Schizophrenia Cognition Rating Scale SCoRS: interview based assessment and its relationship to cognition, real world functioning and functional capacity. *Am J Psychiatry*. 2006;163:426–432.
18. Bowie CR, Twamley EW, Anderson H, Halpern B, Patterson TL, Harvey PD. Self-assessment of functional status in schizophrenia. *J Psychiatr Res*. 2007;41:1012–1018.
19. Marder SR, Fenton WS. Measurement and Treatment Research to Improve Cognition in Schizophrenia: NIMH MATRICS initiative to support the development of agents for improving cognition in schizophrenia. *Schizophr Res*. 2004;72:5–9.
20. Heinrichs DW, Hanlon TE, Carpenter WTJ. The Quality of Life Scale: an instrument for rating the schizophrenia deficit syndrome. *Schizophr Bull*. 1984;10:388–396.
21. Schneider LC, Struening EL. SLOF: a behavioral rating scale for assessing the mentally ill. *Soc Work Res Abstr*. 1983;19:9–21.
22. Jaeger J, Berns S, Czobar P. The Multidimensional Scale of Independent Functioning: A new instrument for measuring functional disability in psychiatric populations. *Schizophr Bull*. 2003;29:153–167.
23. Birchwood M, Smith J, Cochrane R, Wetton S, Copestake S. The Social Functioning Scale: the development and validation of a new scale of social adjustment for use in family intervention programmes with schizophrenic patients. *Br J Psychiatry*. 1990;157:853–859.
24. Schooler NR, Hogarty GE, Weissman MD. Social Adjustment Scale II (SAS). In: Hargreaves WP, Attkisson CC, Sorenson JE, eds. *Resource Materials for Community Mental Health Program Evaluations.* Rockville, MD: US Department of Health, Education and Welfare; 1979:290–302.
25. Wykes T, Stuart E. The measurement of social behavior in psychiatric patients: and assessment of the reliability and validity of the SBS schedule. *Br J Psychiatry*. 1986;148:1–11.
26. Barker S, Barron N, McFarlane B, Bigelow DA. A community ability scale for chronically mentally ill consumers: part I, reliability and validity. *Community Ment Health J*. 1994;30:363–379.
27. Barker S, Barron N, McFarlane B, Bigelow DA. *Multnomah Community Ability Scale: Users Manual.* Portland, OR: Western Mental Health Research Center, Oregon Health Sciences University; 1994.
28. Dickerson FB, Origoni AE, Pater A, Friedman BK, Kordonski WM. An expanded version of the Multnomah Community Ability Scale: anchors and interview probes for the assessment of adults with serious mental illness. *Community Ment Health J*. 2003;39:131–137.
29. Rosen A, Hadzi-Pavlovic D, Parker G. The Life Skills Profile: a measure assessing function and disability in schizophrenia. *Schizophr Bull*. 1989;15:325–337.
30. Wallace CJ, Liberman RP, Tauber R, Wallace J. The Independent Living Skills Survey: a comprehensive measure of the community functioning of severely and persistently mentally ill individuals. *Schizophr Bull*. 2000;26:631–658.
31. Menditto AA, Wallace CJ, Liberman RP, VanderWal J, Jones NT, Stuve P. Functional assessment of independent living skills. *Psychiatr Rehabil Skills*. 1999;3:200–219.
32. Goodman SH, Sewell DR, Cooley EL, Leavitt N. Assessing levels of adaptive functioning: The Role Functioning Scale. *Community Ment Health J*. 1993;29:119–131.