SYMPTOM REPORT AND DIAGNOSTIC PROPERTIES OF THE NEUROPSYCHIATRIC QUESTIONNAIRE

C. Thomas Gualtieri

Dr Gualtieri
NC Neuropsychiatry
400 Franklin Square
1829 East Franklin Street
Chapel Hill NC 27514
919 933 2000 x 106
919 933 2830 fax
tg@ncneuropsych.com

Acknowledgements, Disclosures

The Neuropsychiatric Questionnaire was developed at The North Carolina Neuropsychiatry Clinics in Chapel Hill, Charlotte and Raleigh by Dr. Gualtieri. The NPQ is not a commercial instrument and is freely available on the internet at www.ncneuropsych.com. The research was supported by North Carolina Neuropsychiatry, PA, in Chapel Hill and Charlotte. No external support was sought or received on behalf of this research.

The author thanks Ami Claxton, MS, PhD for her editing assistance in the development of this paper.

ABSTRACT

Background: A problem in psychiatry is the reliance on subjective data, especially self-reported symptoms. In this paper, we investigate the diagnostic validity of the Neuropsychiatric Questionnaire (NPQ) in a large group of patients from a private neuropsychiatric practice.

Method: Patients over age 18 with generalized anxiety disorder (GAD), major depressive disorder (MDD), attention deficit disorder (ADHD), bipolar disorder (BPAD), and normal subjects were included in the analysis (N=1127). Inter-group comparisons were made by MANOVA, controlling for age, gender, education and computer familiarity with the criterion for significance at $P < 0.01$. Analysis of pairwise group differences was by one-way ANOVA with Bonferroni correction. Effect sizes were measured by Cohen's d. The NPQ consists of 20 symptoms scale and four factors: cognitive, mania, somatic and anxiety-depression, which combine to make up a symptom load scale (SLS).

Results: Patient groups differed in the makeup of relative contribution of the four factors to the SLS. When compared to normal subjects (N=45), those in the with (GAD, MDD, ADHD, BPAD), When the four diagnostic groups were compared in pairwise fashion to normal subjects (N=45), patient-reported symptoms occurred in the expected directions, but the effect sizes for the cited differences were, on average, small to moderate. In the discriminant validity analysis, the lowest Wilks' Lambda was 0.622 and patients were correctly classified by the NPQ ranging from 65-79% of the time.

Conclusion: The NPQ, while being a useful tool for the clinician, deserves a diagnostic weight no higher than 24%. The clinical history, family history and examination deserve more weight. This tool may be more appropriate for tracking symptoms over time than for initial diagnosis. The average lambda score in Table 4 is 0.76. If that metric is at all meaningful, then what it means is that patient self-report contributes 24% to diagnostic discrimination and 76% comes from other sources.

INTRODUCTION

A problem in psychiatry is the reliance on subjective data, especially self-reported symptoms. Proper diagnosis is, of course, not simply a function of patient self-report. The clinical history,family history and mental state examination are essential components of the diagnostic process. However, clinicians rely to an increasing degree on what patients tell them about how they feel, especially in the present climate, where only 30 minutes may be afforded for an initial evaluation and 15 minutes for a follow-up "med check."

One way to systematize the collection of self-reported symptoms and give subjective data a quantitative dimension is the use of rating scales (RS) and symptom checklists (SCL). Theoretically, RS allow for uniformity of assessment and comparability of results across patients and in different clinical sites, which is why they are used in psychiatric research and clinical trials. They turn subjective data about patients' symptoms into quantitative data with at least a measure of objectivity. In clinical practice, the use of rating scales are said to have positive ramifications on patient outcome and quality assurance review.[1 2] Support for ongoing treatment by third-party carriers, for example, is increasingly posited on the demonstration of treatment efficacy. Quality review is often based of patient's subjective response to a clinical interaction.

If it were possible to capture patients' self-report by using a broad-spectrum SCL, would that have an impact on the diagnostic process? That is a question we addressed by examining the diagnostic relevance of a computer-administered SCL to a large group of patients with four common neuropsychiatric disorders.

METHODS

THE NEUROPSYCH QUESTIONNAIRE

The first description of the Neuropsychiatric Questionnaire (NPQ) was published in 2007 in an open-source internet journal.[3] It described item selection, the generation of symptom scales, the test-retest reliability of the test and its sensitivity to treatment. In a second paper, we described the factor structure of the NPQ, its correlation with commonly used rating scales, and concordance between reports from patients and spouses (Gualtieri, 2013, under review).

The NPQ is a broad-spectrum symptom checklist appropriate for use in patients with neuropsychiatric disorders. The adult version of the NPQ consists of 207 questions about common symptoms of neuropsychiatric disorders (Gualtieri, 2007). Each item is scored as "not a problem" (0), a "mild problem" (1), a "moderate problem (2) or a "severe problem" (3). The 2007 analysis established that the 207 items clustered into 20 symptom scales. For example, 22 of the 207 items addressed the problem of memory impairment; the patient's scores on these 22 items are averaged and multiplied by 100. As such, a memory symptom scale score is generated; the highest score someone might achieve on a symptom scale is 300, which would indicate that the patient marked every item in the relevant scale as "a severe problem." The lowest score one might get is zero, if every item in the scale was scored as "not a problem." The 20 symptom scales in the NPQ are: attention, hyperactivity-impulsivity, learning problems, memory, anxiety, panic, agoraphobia, obsessions and compulsions, social anxiety, depression, mood instability, mania, aggression, psychosis, somatization, fatigue, sleep, suicide, pain and substance abuse.

The 20 symptom scales in the NPQ refer to symptom clusters and not to DSM or ICD diagnoses. This was a considered decision because the NPQ was designed to be a measurement instrument, not a diagnostic tool. For example, the depression scale asks questions about depression itself, not anxiety, fatigue or sleep difficulties,

which are addressed separately in the corresponding relevant scales. In this way, the NPQ is different from conventional rating scales, like the Hamilton Depression Rating Scale, which contains a number of anxiety-related items, or the Conners Parent-Teacher Questionnaire, which includes items related to inattention, hyperactivity-impulsivity and emotional instability.

Of the 20 scales, 17 load with four factors: a cognitive factor (CF), an anxiety-depression factor (ADF), a mania factor (MF) and a somatic factor (SF). The remaining 3 scales – suicide, sleep and substance abuse – do not load with any of the four factors. Factor scores are calculated by averaging the symptom scales within the factor, and range from 0 to 300. A symptom load scale (SLS) is generated by summing the factor scores; the SLS may range from 0 to 1200.

TABLE 1. TWENTY SCALES IN THE ADULT NPQ

Symptom Scale	Factor
Inattention	Cognitive
Learning problems	Cognitive
Memory problems	Cognitive
Anxiety	Anxiety-Depression
Panic	Anxiety-Depression
Agoraphobia	Anxiety-Depression
Obsessions, compulsions	Anxiety-Depression
Social anxiety	Anxiety-Depression
Depression	Anxiety-Depression
Hyperactivity-Impulsivity	Mania
Mood instability	Mania
Mania	Mania
Aggression	Mania
Psychosis	Mania
Pain	Somatic
Somatization	Somatic
Fatigue	Somatic
Disordered sleep	*
Suicide	*
Substance abuse	*

*Does not load with any of the four factors

SUBJECTS

The data in this report are from adult patients who were evaluated at the Neuropsychiatry Clinics in Raleigh, Chapel Hill and Charlotte, North Carolina (NCNC) over a five year period (2006-2011). Demographic data, age, race, gender, level of education and self-reported computer familiarity are collected when the NPQ is administered. At the time of administration, patients give written informed consent to the use of their de-identified clinical data, including the NPQ, for the purposes of research and program evaluation, and patients may rescind their consent at any time.

In the NCNC database containing more than 16,000 NPQ files, the records of 1127 subjects were selected who met the following criteria: over age 18, diagnoses of generalized anxiety disorder (GAD), major depression (MDD), attention deficit/hyperactivity disorder (ADHD) or bipolar disorder, depressed (BPAD); 45 normal subjects from a previous validation study of the NPQ were also included. It is important to note that diagnoses were based on a comprehensive examination, of which the NPQ battery was only a part; diagnoses were not made simply on

the basis of the NPQ scores. Diagnoses were made on the basis of DSM-IV TR criteria, where applicable and affirmed by review by a research psychiatrist (CTG). Bipolar patients were either BPAD-I or BPAD-II, but none were manic at the time of evaluation. Every attempt was made to eliminate patients with multiple psychiatric diagnoses, active medical conditions, neurological conditions or cognitive disorders.

When a patient is evaluated at one of the clinics, he or she is routinely administered the NPQ, along with other rating scales, cognitive tests and validity measures, as appropriate. The NPQ data are automatically uploaded into a central database, which is maintained under secure conditions and available only to selected clinicians in the practice.

ANALYSIS

Inter-group comparisons were made by MANOVA, controlling for age, gender, education and computer familiarity with the criterion for significance at $P < 0.01$. Analysis of pairwise group differences was by one-way ANOVA with Bonferroni correction. Effect sizes were measured by Cohen's d.

RESULTS

Table 1 presents the diagnoses, numbers of subjects in each category, age, race, gender and level of education. The youngest subject was 18 years old. The oldest, a normal subject, was 79. The population of subjects was predominantly white (78%), 14% black and 8% Hispanic, Asian or Native American. The five groups differed significantly in age, level of education, gender and computer familiarity; these variables were controlled in subsequent analyses.

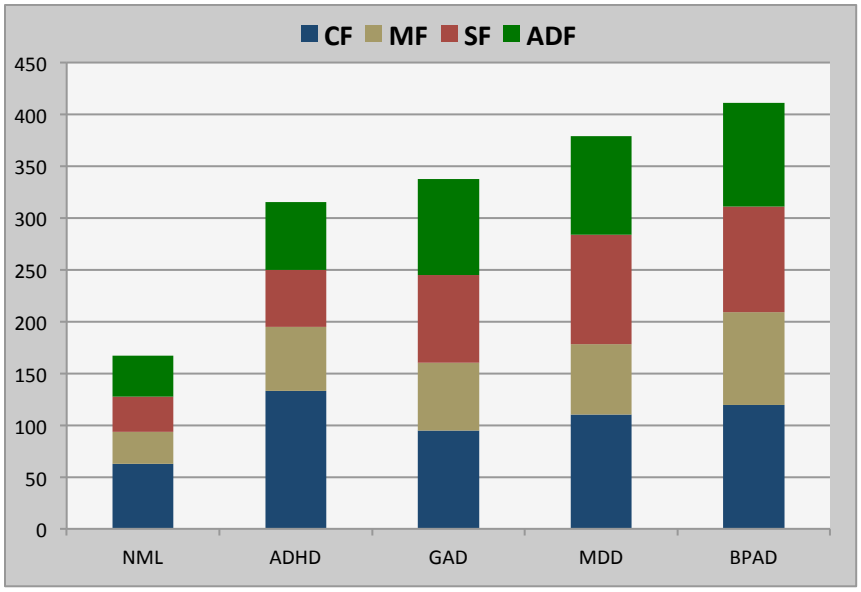
TABLE 2. DEMOGRAPHIC CHARACTERISTICS OF THE CLINICAL POPULATION

	NML		GAD		MDD		ADHD		BPAD		ANOVA P <
N	45		140		345		444		153		
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
AGE	35.6	10.9	36.7	12.7	40.8	14.4	30.6	12	38.8	12.4	.000
EDUC	15.7	1.9	14.5	2.5	14.4	2.3	14.7	2.2	14.1	2.3	.005
COMP FAMILIARITY	2.9	0.3	2.8	0.5	2.7	0.6	2.8	0.4	2.7	0.5	.000
SEX % MALE	0.33		0.38		0.34		0.51		0.26		.000
RACE % WHITE	0.69		0.84		0.79		0.79		0.75		.389

*Computer familiarity is self-reported as 1, "none," 2, "some," or 3, "frequent."

The structure of the NPQ permits a graphic display of patient data, as shown in Figure 1. The SLS is a measure of the patient's total symptom burden, the sum of the four factor scores (min = 0, max = 1200). Visual inspection of the figure suggests that normal subjects reported more cognitive symptoms compared to somatic, manic or anxiety-depression symptoms. ADHD patients tend to show more cognitive symptoms compared to patients with GAD, MDD or BPAD. BPAD patients did not show appreciably more symptoms in the mania factor than the GAD or MDD patients, presumably because most of them were in the depressed state and none were overtly manic.

FIGURE 1. THE SLS AND FACTOR SCORES IN FIVE GROUPS



y-axis: Symptom Load Scale (SLS)

x-axis: NML (normal control), ADHD (attention deficit hyperactivity disorder), GAD (generalized anxiety disorder), MDD (major depressive disorder), BPAD (bipolar disorder)

When the five groups are compared by MANOVA, they differed significantly in the SLS, the four factor scales and in nineteen of the twenty symptom scales. The only symptom scale that did not score differently was substance abuse; all of the groups scored very low in that scale. Analysis with Bonferroni correction indicated that many of the comparisons were not statistically significant. Table 2 presents the effect sizes (Cohen's d) for comparisons of the four patient groups to normal Ss that were significant at the level of $P < 0.01$. The patient groups all had higher symptom load scores. GAD patients, compared to normal Ss, had higher scores in the Mania, Somatic and Anxiety-depression factors and their associated symptom scales. ADHD patients were higher in the cognitive and mania factors. MDD and BPAD patients were higher in all of the symptom scores, save social anxiety and substance abuse.

The effect sizes of the differences were for the most part large (Cohen's $d > 0.7$).

TABLE 2. NORMAL SUBJECTS COMPARED TO FOUR GROUPS OF PSYCHIATRIC PATIENTS

NML Ss COMPARED TO:				
	GAD	MDD	ADHD	BPAD
SLS	0.86	1.01	1.01	1.09
CF		0.69	0.69	0.83
MF	0.75	0.77	0.77	1.04
SF	0.87	1.06		1.00
ADF	0.95	0.97		0.95
ATT		0.67	0.67	0.83
LPX		0.63	0.63	0.74
MEM		0.64	0.64	0.75
ANX	1.13	0.96	0.96	0.96
PANIC	0.94	0.78		0.81
AGORA	0.57	0.68		0.69
OC		0.55		0.66
SAD				
DEP	0.86	1.27	1.27	1.11
HIP	0.80	0.68	0.68	1.00
MS	0.68	0.80		1.02
MANIA	0.78	0.71	0.71	0.85
AGG				0.64
PSYCH		0.58		0.79
SOMA	0.73	0.80		0.75
FTG	0.83	1.12		1.00
PAIN	0.69	0.84		0.79
SLEEP	0.87	0.83		0.92
SUI		0.69		0.83
SA				

When the four diagnostic groups were compared in pairwise fashion, patient-reported symptoms occurred in the expected directions. Compared to GAD patients, MDD patients reported higher levels of depression and fatigue. GAD patients reported more somatic and anxiety depression symptoms than ADHD patients, and ADHD patients reported more cognitive symptoms than GAD patients. BPAD patients reported more manic symptoms than GAD patients, and had a higher symptom load. MDD patients reported a higher symptom load than ADHD patients and reported more somatic and anxiety-depression symptoms. BPAD patients reported more manic symptoms than MDD patients. They reported a higher symptom load than ADHD patients, but the two groups were not notably different in their level of cognitive symptoms. The effect sizes for the cited differences were, on average, “small” or “moderate.”

TABLE 3. PAIRWISE COMPARISON OF FOUR PATIENT GROUPS

	GAD			DEP		ADHD
	MDD	ADHD	BPAD	ADHD	BPAD	BPAD
SLS			0.37	0.35		0.55
CF		0.59		0.35		
MF			0.47		0.42	0.61
SF	0.33	0.57		0.82		0.83
ADF		0.54		0.56		0.65
ATT		0.70	0.42	0.47		
LPX		0.55		0.39		
MEM		0.36				
ANX		0.60		0.35		0.47
PANIC		0.75		0.51		0.63
AGORA		0.39		0.49		0.54
OC				0.25		0.45
SAD						0.33
DEP	0.52		0.44	0.81		0.76
HIP					0.39	
MS			0.46	0.35	0.33	0.69
MANIA			0.40		0.37	0.49
AGG			0.48		0.40	0.50
PSYCH			0.35	0.28		0.59
SOMA		0.56		0.64		0.65
FTG	0.42	0.40		0.77		0.73
PAIN		0.57		0.75		0.75
SLEEP		0.56		0.48		0.63
SUI			0.53	0.51		0.76
SA						

Table 4 shows the results of discriminant function analysis in a series of pairwise comparisons among the four patient groups. The symptom scales that discriminate are consistent with the diagnostic definitions of the various disorders

TABLE 4. SUMMARY DATA FROM DISCRIMINANT FUNCTION ANALYSIS

COMPARISON		Eigen-value	Wilks' Lambda	Correctly Classified	SCALES THAT DISCRIMINATE	
DX1	DX2				DX1	DX2
GAD	MDD	0.247	0.802	0.73	• Anxiety • Social Anxiety	• Depression
GAD	ADHD	0.424	0.702	0.79	• Panic • Anxiety • Obsessions, Compulsions • Sleep	• Inattention • Memory • Hyperactivity-Impulsivity
GAD	BPAD	0.289	0.776	0.73	• Anxiety	• Suicide • Mood Instability • Mania • Inattention
MDD	ADHD	0.608	0.622	0.78	• Depression • Pain • Aggression • Fatigue	• Inattention • Memory Problems • Learning Problems

					• Social Anxiety	
MDD	BPAD	0.115	0.897	0.65	• Depression	• Suicide • Mood Instability • Mania
ADHD	BPAD	0.352	0.74	0.75	• Learning Problems • Inattention • Memory Problems	• Suicide • Pain • Sleep • Mania

DX1 = Diagnosis 1; DX2 = Diagnosis 2

DISCUSSION

The NPQ, a broad-spectrum SCL, behaves in an orderly manner. The relative order of an overall symptom load scale, the four factor scores and at least 19 of the 20 symptom scales reveals a coherent relationship to four common neuropsychiatric diagnoses. The scores that differ significantly when patients in the four groups are compared to normal subjects, and when the four diagnostic groups are compared to one another are reflective, at least to a degree, with the known symptoms of the conditions. In a sense, these comparisons may be interpreted as support for the criterion validity of the NPQ.

The advantage of patient-reported symptom checklists is primarily one of efficiency: clinician time is not required, and one can systematically inquire after a broad spectrum of pertinent symptoms. Since the NPQ is internet-based, it can be administered to patients' family members or other acquaintances, giving confirmatory information about the patient's condition.

The disadvantage is, ostensibly, reliability: patients may interpret questions in idiosyncratic ways, and response patterns may differ in patients from different ethnic or socio-economic groups [4]. Patients with severe disorders may have limited insight. Some patients may be inclined to color their responses for one reason or another.

The major problem of the NPQ, however, is its diagnostic utility, or lack thereof. The data in Table 4 indicate symptom scales that discriminate one diagnosis from another. The results are intuitive; patients with GAD, for example, are identified by the anxiety and social anxiety scales, patients with MDD by the depression scale. None of the six pairwise analyses, however, generate eigenvalues higher than 0.608 or lambda's less than 0.622, and only 65-79% of patients are correctly classified on the basis of their NPQ scores alone.

In the course of this investigation, the author examined other ways to examine the relevance of the NPQ to psychiatric diagnosis. For example, we standardized the scale scores for patients based on the scores of normal subjects.[3] We examined every symptom score as a proportion of the SLS. We looked for patterns in the relative proportion of factor scores, controlling for difference from normal. We developed indices based on the discriminating symptom scores and ran ROC analyses. None of these manipulations changed the essential result of the investigation: depending on one's point of view, the results are encouraging or disappointing. High scores in symptom scales and scale factors do have a measure of diagnostic relevance. But no one can pretend that the NPQ is a diagnostic instrument. The average lambda score in Table 4 is 0.76. If that metric is at all meaningful, then what it means is that patient self-report contributes 24% to diagnostic discrimination and 76% comes from other sources.

In evaluating the salience of the symptom scales to patient diagnosis, we elected to make pairwise comparisons between selected diagnoses. In most of the examples we presented, the F statistic was of limited use; virtually all of the differences were statistically significant. Judging the magnitude of the differences by estimating effect sizes identified scales where meaningful differences occurred. When we did that, the scales that prove to be important are usually the ones one would expect. When patients with anxiety, for example, were compared to patients with depression, the panic and anxiety scales are at one pole and depression, fatigue, suicide, pain and memory are at the other. Applying a more rigorous statistic to the analysis, we evaluated conducted stepwise discriminant function analysis. Not surprisingly, the results were similar. We report data from the latter analysis, however, because the degree to which the symptom scores contribute to the discriminant function can be expressed by Wilks' Lambda. As it happens, few of the lambdas are low.

We are satisfied, however, that the relative weakness of the NPQ as a diagnostic instrument is not because of any inherent weakness in the questionnaire; the psychometric properties of the NPQ are sound. The weakness of a SCL like the NPQ is not related to unreliable reporting. We have previously reported good test-retest reliability for the NPQ and good correlations between patients and spouses (r = 0.29-0.80) (Gualtieri, under review, 2013). The correlations between the relevant NPQ symptom scales and five commonly used clinical rating scales, including clinician administered RS like the Ham-D and Ham-A, is moderate.[3] (Gualtieri, under review, 2013) The correlations are consistent with published results of clinician RS compared to patient self-report RS. For example, in a review of seven studies comparing the HamD to the BDI, the correlations ranged from 0.21 to 0.82 [5] and in nine similar studies, correlations between the HamD and the Zung Self-rating scale Depression Scale ranged from 0.22 to 0.95 (Hedlund et al, 1979; McDowell, et al 1996).[6 7] The reliability of the NPQ is more-or-less equivalent to published data for other scales.

If we failed to establish discriminant validity, it could be because of flaws in the research. The NPQ was part of the diagnostic evaluation, so evaluating its diagnostic salience is a bit circular . The diagnoses themselves were given by clinicians, and although they were reviewed by a research psychiatrist they were not arrived at in the fashion that is current at research centers. Every measure was taken to minimize diagnosis heterogeneity or comorbidity; that is, every measure that is feasible in a clinical setting. On the other hand, the issue of diagnostic discrimination is, arguably, more important in a routine clinical setting because of time-constraints and the reliance clinicians have, under such circumstances, on patient report.

The likely conclusion one can draw is that rating scales in general, and the patient-reported NPQ in particular, are insufficient grounds for making a diagnostic discrimination. They may be useful as screening instruments or as measures of change over time, but not as diagnostic tools. This, of course, is a truism. In psychiatric research, when systematic diagnosis is essential, instruments like the Structured Clinical Interview for DSM Disorders (SCID) are used. They are the antithesis of a SCL, but they may take 45-90 minutes to administer and are unsuitable for clinical practice.

There is a lesson to be gleaned from this for the day-to-day practice of clinical psychiatry. In an era of limited reimbursement for psychiatric evaluation, it is tempting to rely on rating scales or symptom checklists. As screening measures they are useful, but they are not diagnostic instruments. Patient self-report deserves no higher weight than 24%. The clinical history, correlative data, family history and mental state examination deserve more weight.

The NPQ is a simple instrument. It uses a PC to ask the patient or a knowledgeable informant about patients' symptoms, perceived within a designated time frame. But it is no more than what it appears: a SCL that is quick, cheap and easy to use. If it is not comprehensive, at least it covers a lot of ground. It is necessarily limited because it relies on an informant's ability, or his inclination, to disclose accurate information. Thus, the report generated by the NPQ includes this *caveat*:

A high score means that the patient is reporting more symptoms of greater intensity. It doesn't necessarily mean that the patient has a particular condition; just that he or she (or their spouse, parent or caregiver) are saying that he or she has a lot of symptoms in a particular area. Conversely, a low score simply means that the patient (or caregiver) is not reporting symptoms associated with a particular condition, at least during the period of time specified. It does not mean that the patient does not have a condition. Just as some people over-state their problems, others tend to under-state their problems. The NPQ is not a diagnostic instrument. The results it generates are meant to be interpreted by an experienced clinician in the course of a clinical examination.

REFERENCES

Berzon, R, R D Hays, and S A Shumaker. "International Use, Application and Performance of Health-Related Quality of Life Instruments." *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 2, no. 6 (December 1993): 367–368.

Dowd, Jennifer Beam, and Megan Todd. "Does Self-Reported Health Bias the Measurement of Health Inequalities in U.S. Adults? Evidence Using Anchoring Vignettes from the Health and Retirement Study." *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences* 66, no. 4 (July 2011): 478–489. doi:10.1093/geronb/gbr050.

Gualtieri, CT. "An Internet-Based Symptom Questionnaire That Is Reliable, Valid, and Available to Psychiatrists, Neurologists, and Psychologists." *MedGenMed: Medscape General Medicine* 9, no. 4 (2007): 3.

Hedlund, JL, and BW Vieweg. "The Hamilton Rating Scale for Depression: A Comprehensive Review." *J Operational Psychiatry* 10 (1979): 149–165.

Kesselheim, Aaron S, Timothy G Ferris, and David M Studdert. "Will Physician-Level Measures of Clinical Performance Be Used in Medical Malpractice Litigation?" *JAMA: The Journal of the American Medical Association* 295, no. 15 (April 19, 2006): 1831–1834. doi:10.1001/jama.295.15.1831.

McDowell, I, and C Newell. *Measuring Health. A Gude to Rating Scales and Questionnaires*. 2nd ed. New York: Oxford Univ Press, 1996.

Slade, Mike, Paul McCrone, Elizabeth Kuipers, Morven Leese, Sharon Cahill, Alberto Parabiaghi, Stefan Priebe, and Graham Thornicroft. "Use of Standardised Outcome Measures in Adult Mental Health Services: Randomised Controlled Trial." *The British Journal of Psychiatry: The Journal of Mental Science* 189 (October 2006): 330–336. doi:10.1192/bjp.bp.105.015412.