Assessing the Sources of Unreliability (Rater, Subject, Time-Point) in a Failed Clinical Trial Using Items of the Positive and Negative Syndrome Scale (PANSS)

Anzalee Khan, PhD,† William Christian Yavorsky, PhD,‡ Stacy Liechti, PhD,† Guillermo DiClemente, PhD,‡ Brian Rothman, PhD,† Mark Opler, PhD, MPH,†§ Ashleigh DeFries, MA,// and Sofija Jovic, PhD†*

Background: Considering the increasing attention to the study of failed clinical trials, the goal of this study was to identify the sources of unreliability in a failed clinical trial by assessing scores on the Positive and Negative Syndrome Scale (PANSS).

Methods: This study is a substudy from a failed phase 2 double-blind, placebo-controlled trial of schizophrenia. Using the generalizability theory, this substudy assesses reliability on 3 conditions: raters, time points (PANSS evaluations, 1 week apart), subjects for 3 groups (placebo responders, placebo nonresponders, and treatment group).

Results: The placebo response rate was 40.07% (32/71). For all PANSS positive symptom items, the most variability was for raters (range, 33%–72%) for the placebo responders, 31% to 68% for the placebo nonresponders, and 29% to 60% for the treatment group. The variability of the interaction of rater and time point was the second source of unreliability, with an average of 12.28% compared to 12.00% for the placebo nonresponders and 10.00% for the treatment group. All items of the negative symptom subscale showed the most percent variability for raters, for all groups. For general psychopathology items (except preoccupation), raters accounted for the most variability in the scores for placebo responders with an average of 51.00% across items. A similar pattern was observed for the placebo nonresponders and for the treatment group; for the treatment group, the interaction between rater and time point accounted for the most variability for somatic concern and anxiety.

Conclusions: Results confirm the efficacy of applying the generalizability theory to the estimation of reliability to identify a source of unreliability and provide evidence for the relationship between low reliability and failed trials. Findings can be used to guide data monitoring, rater training, and identification of PANSS items, which may require supplementary training.

Key Words: generalizability theory, PANSS, reliability, rater training, schizophrenia

(*J Clin Psychopharmacol* 2013;33: 109–117)

Failed trials are a problem for the development of pharmacological treatments. A trial is considered failed when the active treatment does not differentiate from placebo. Contributing factors may include the following: escalating placebo response rates, dosing regimens, low sensitivity in the efficacy measures, and inconsistency in rating scores, especially in multicenter trials. Interim monitoring of assessment tools, clinical outcomes, and measurement errors is an important tool for early decision making.[1,2] In psychiatry, double-blind, placebo-controlled trials that have failed to confirm the superiority of a drug over a placebo are prevalent.[3] Otto and Nierenberg[4] indicate that to assume that a trial has failed owing to unsuccessful assay sensitivity misrepresents the scientific practice, especially when the trial is judged solely by the results and not by the trial design, rater variability, or in-study data monitoring techniques. Among the contributing factors to the problem identified thus far, placebo response is one of the most relevant.[3] Placebo response is the improvement in the clinical condition of patients who are assigned randomly to the placebo group.[5] In the development of assessment instruments for psychopathology, there has been a tendency to use standardized rating procedures using trained clinicians or video-based administration and scoring. An integrated symptom assessment approach, involving subject, visits, and raters, would be of great value across the spectrum of schizophrenia.

The 30-item Positive and Negative Syndrome Scale (PANSS)[6] is rated by a clinician and is commonly used in clinical trials to assess the severity level of psychopathology identified with schizophrenia and related disorders. Before rating the PANSS, a clinician participates in comprehensive training, and clinicians' scores are compared to a criterion standard rating.[7]

The psychometric properties of the PANSS have been extensively examined through classical test theory and, more recently, by latent trait modeling.[8,9] The most familiar methods of examining reliability involve evaluating the consistency of measurement over repeated assessments by the same rater (test-retest reliability), across different raters (interrater reliability), and across items (internal consistency). For each of these methods, the fundamental theory is that the level of psychopathology is a stable, fixed trait and any variability is as a result of measurement error.[10] The generalizability theory (G theory) is an alternative to the standard reliability approaches, as it accounts for the differences in measurement conditions and increases generalizability by dividing the error variance into separate sources.[11] To classify singular sources of error, studies may be designed purposely to examine the reliability of clinician scores by various features (facets).[12] A facet is a set of conditions that, because of the differences among the conditions, may contribute to the observed variability among the scores. For example, in the assessment of psychopathology, the person doing the assessment (rater facet), the types of symptoms experienced (subject facet), and the fluctuating course of the disease process (time-points facet) could likely influence the variability of symptom assessments. In lieu of merging numerous sources of measurement variance by a common error term,

From the *Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY; †ProPhase LLC, New York, NY; ‡CROnos Clinical Consulting Services, Hamilton Township, NJ; §New York University School of Medicine, New York, NY; and ||Johns Hopkins University, Baltimore, MD.
Received October 5, 2011; accepted after revision April 26, 2012.
Reprints: Anzalee Khan, PhD, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY; ProPhase LLC, 3 Park Ave, 37th Floor, New York, NY 10016 (e-mail: AKhan@nki.rfmh.org).
This project was funded by ProPhase, LLC, NY, Investigator-Initiated Trial.

ISSN: 0271-0749
DOI: 10.1097/JCP.0b013e3182776ebe

the G theory facilitates the separating of error variance by its source.[11,13] Owing to the classification and inspection of sources of error, the accuracy of scores can be improved.[14] Studies have applied the G theory to rating scales[15] and assessments[16] to identify sources of variation and their impact on reliability.

Using the principles of the G theory, we assessed the extent to which each facet (ie, raters, subjects, and time-points [visits]) contributed to the variability (and inconsistency) in scores of the PANSS for subjects identified as placebo responders, placebo nonresponders, and those assigned to treatment.

MATERIALS AND METHODS

Design

Data Source

This study was designed as a substudy from a failed phase 2 multicenter, randomized, double-blind, placebo-controlled, parallel-group, flexible-dose, 10-week study into the safety and efficacy of an oral antipsychotic (partial agonism at dopamine D2/D3 receptors, with preferential binding to D3 receptors, and partial agonism at serotonin 5-HT1A receptors) and placebo, for schizophrenia, from November 2006 to August 2007 across 41 centers (United States).

The original study included male and female inpatients, 18 to 65 years of age, meeting *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision* criteria for schizophrenia and having a total PANSS score of 70 or higher and 140 or lower, with a score of 4 or higher (moderate) on two of the following items: delusions (P1), conceptual disorganization (P2), hallucinatory behavior (P3), and suspiciousness/persecution (P6), where at least one of the items is either delusion or hallucinatory behavior. The duration of treatment was up to a 7-day washout period followed by a 6-week double-blind treatment and a 4-week safety follow-up period (there was a total of 8 treatment visits). Using the principles of the G theory,[16] this substudy is a 3-group design (placebo responders, placebo nonresponders, and treatment group) to assess the reliability of symptom ratings according to 3 facets: (1) raters (trained clinicians), (2) time points (8 PANSS evaluations, 1 week apart), and (3) subjects.

Based on a priori criteria of 20% improvement in PANSS total scores, subjects with 20% or greater improvement (at any visit) in PANSS total score were defined as responders. The cutoff of 20% improvement was based on research showing that minimal improvement on the Clinical Global Impression—Severity of Illness[17] is associated with 23% reduction in the PANSS total score at 2 weeks[18] and is consistent with studies assessing the relationship between early response and nonresponse.[19]

Sample

Of the 392 subjects, 262 subjects were assigned to treatment and 130 subjects were assigned to placebo. For the treatment group, 141 subjects completed all visits; and for the placebo group, 71 subjects completed all visits (8 visits). The mean ± SD age of subjects was 41.15 ± 9.90 years. Most of the placebo and treatment group subjects were African American, 76.16% (n = 49) and 61.10% (n = 160), respectively. For the placebo group, 84.51% (n = 60) were men; whereas for the treatment group, 79.40% (n = 208) were men. The PANSS scores ranged from 75 to 117 and 73 to 134 at baseline for the placebo and treatment groups, respectively. There were 59 raters used in the entire study. The placebo response rate observed in this study was 40.07% (32/71). For the treatment group, 60.28% (85/141) were treatment responders. For this substudy, only patients who completed all 8 visits were included in the G theory analysis. Because this study does not aim to test a hypothesis about treatment effects, conventional sample size calculations are not relevant. Additionally, the G theory can use small sample sizes for reliability estimates.[20]

Measure and Rater Training

The PANSS is a 30-item rating scale to assess the severity schizophrenic psychopathology. All items are rated on a 7-point scale (1, absent; 7, extreme) and include 7 positive subscale items (P1-P7), 7 negative subscale items (N1-N7), and 16 general psychopathology subscale items (G1-G16). A semi-structured clinical interview, the SCI-PANSS,[21] was used to guide raters through specific questions to rate symptom severity. Items are summed for the 3 subscales and total score.

Each PANSS rater obtained rater training and certification. Rater certification included achieving an interrater reliability (Cronbach α, 0.80) with "Expert Consensus Gold Standard PANSS" scores, derived from one rating by qualified psychiatrists/psychologists who completed adequate rater training. Rater training was completed before the initiation of the study and was not repeated throughout the course of the study.

Statistical Analysis

The numbers of placebo and treatment subjects randomized in the study who completed the study and those who discontinued the study were tabulated by counts and proportions overall. Interrater reliability was tested based on intraclass correlation coefficient (ICC) for all patients (n = 392), for the placebo group (n = 130), and for active treatment group (n = 262). We classified ICC greater than 0.75 as excellent agreement and less than 0.4 as poor agreement.[22]

The G theory was used to assess multiple sources of error variation among groups.[12] For the PANSS ratings, raters, visits, and subjects are 3 error sources. As in analysis of variance, the observed score (rating given to the subject) is divided up into the grand mean and components of the main and interaction effects in addition to random error. Each of these components (apart from the grand mean) has a variance component. Therefore, the variance of each PANSS score is equal to the sum of the variance components. Using these variance components attributable to each independent source of error variance (or facet) and interactions among the facets allows clinicians to make decisions about how to minimize the effect of error variance. Thus, the G theory leads into decision (D) studies. In a particular, PANSS subscale, a D study, can help the researcher and clinician in making decisions about the optional number of visits that give dependable ratings. Decision studies allow the researcher to estimate how reliability coefficients improve if different aspects of measurements are altered. For this study, the G coefficients used are Ep^2 and ϕ. Ep^2 is the equivalent of a reliability coefficient (such as, Cronbach α) and denotes the consistency in the relative scores of the subjects and is used for relative reliability, which is only concerned with ranking individuals and not with item difficulty.[12] The ϕ is the index of dependability and is used for absolute reliability (important when making criterion-referenced decisions) and includes all variance components except the object of measurement.[12] Similar to conventional reliability, the values of Ep^2 and ϕ range from 0 to 1, with higher values reflecting more dependable measurements.

For this G theory analysis, only subjects who completed all visits were used for the analysis (N = 212: placebo responders, 32; placebo nonresponders, 39; and treatment group, 141). Brennan[11] and Cronbach et al[12] recommend calculating the mean square for each effect through analysis of variance and

then equating each source to its expectation (expected mean square). All analyses were performed in SPSS for Windows, Version 17.0.

In the interest of confidentiality, no treatment code information was included in the data, nor was there any exchange of information that might identify the subjects. The study was approved by the institutional review board at participating sites for efficacy analysis and for secondary analysis of existing data.

RESULTS

Interrater Reliability

At baseline, ICC was 0.70 (95% CI, 0.65–0.74) for the entire sample. The interrater reliability for the placebo sample was ICC = 0.68 (95% CI, 0.60–0.76), with the ICC for the placebo responders being 0.66 (95% CI, 0.58–0.74) and the ICC for the placebo nonresponders being 0.70 (95% CI, 0.64–0.78). The ICC at baseline was 0.71 (95% CI%, 0.65–0.75) for the treatment group. At end point, interrater reliability was ICC = 0.77 (95% CI, 0.70–0.81) for the entire sample. The ICC for the placebo sample at end point was 0.75 (95% CI, 0.69–0.80), with the ICC for the placebo responders being 0.71 (95% CI, 0.63–0.75) and that of the placebo nonresponders being 0.72 (95% CI, 0.66–0.80). The ICC at end point was 0.79 (95% CI, 0.70–0.82) for the treatment group.

Source of Measurement Error

Positive Symptoms

The sources of variance attributed for each of the 7 items of the PANSS positive subscale for the placebo responder, placebo nonresponders, and treatment groups are presented in Table 1. For every positive subscale item, raters accounted for the most variability in the given scores, ranging from 33.00% (P7, hostility) to 72.00% (P1, delusions; and P4, excitement), with a mean of 60.71% across the 7 positive symptom items in the placebo responder group; 31.00% (P7) to 68.00% (P3, hallucinatory behavior), with a mean of 57.00% for the placebo nonresponder group; and 29.00% (P7) to 60.00% (P1 and P3 each) with a mean of 51.00% for the treatment group. The next largest source of variance for the placebo responder group was attributed to the interaction between the rater and visit, with a mean of 12.28%, compared to 12.00% for the placebo nonresponders, and 10.00% for the treatment group.

The $E\rho^2$ ranges from poor (P1) to excellent (P7) for all 3 groups, with a mean of 0.727 across items for the placebo responders, 0.662 for the placebo nonresponders, and 0.680 for the treatment group (Table 1). The ϕ ranges from poor (P1) to very good (P7), with a mean of 0.637 across items for the placebo responders, 0.615 for the placebo nonresponders, and 0.647 for the treatment group. For all 3 groups, $E\rho^2$ and ϕ indicate that the PANSS positive subscale item scores of hostility and excitement show good to excellent reliability, whereas grandiosity, hallucinatory behavior, conceptual disorganization, and suspiciousness/persecution are fair; and the ratings of delusions is poor.

Negative Symptoms

For the placebo responder group, rater variability ranged from 54.00% (N6, lack of spontaneity/flow of conversation) to 69.00% (N1, blunted affect; N5, difficulty in abstract thinking). For the placebo nonresponder group, rater variability ranged from 49.00% (N6) to 64.00% (N1). For the treatment group, rater variability ranged from 45.00% (N6) to 56.00% (N1) (Table 2). The next largest source of variance was the interaction between rater and visit, with a mean of 11.00% for the placebo responders, whereas the placebo nonresponder group and the treatment group both had a mean of 9.00%.

The $E\rho^2$ ranges from poor (N4, passive apathetic social withdrawal; N2, emotional withdrawal; N6; N3, poor rapport) to excellent (N5; N1; N7, stereotyped thinking), with a mean of 0.647 across items for the placebo responders, 0.667 for the placebo nonresponders, and 0.675 for the treatment group. The ϕ ranges from poor (N6, N4, N2, and N3) to good (N5 and N1), with a mean of 0.599 across items for the placebo responders, 0.632 for the placebo nonresponders, and 0.635 for the treatment group.

General Psychopathology

For all general psychopathology items (except G15, preoccupation), raters accounted for the most variability in the scores for the placebo responders, ranging from 33.00% (G1, somatic concerns) to 70.00% (G6, depression), with a mean of 51.00% across the 16 items (Table 3). For G15, the interaction between raters and visit accounted for the most variability (49.00%). A similar pattern was observed for the placebo nonresponders and for the treatment group; however, for the treatment group, the interaction between rater and visit also accounted for the most variability for G1; and G2, Anxiety. The $E\rho^2$ averages 0.711 for the placebo responders, 0.740 for the placebo nonresponders, and 0.769 for the treatment group. The ϕ averaged 0.646 for the placebo responders, 0.669 for the placebo nonresponders, and 0.691 for the treatment group. $E\rho^2$ and ϕ indicate that the PANSS general psychopathology symptom items scores of disorientation, depression, motor retardation, mannerisms and posturing, and guilt feelings show good to excellent reliability across all 3 groups.

PANSS Total

The $E\rho^2$ for all 30 items averages 0.687, 0.705, and 0.727 for the placebo responders, placebo nonresponders, and treatment groups, respectively, denoting moderate consistency in the PANSS scores. The treatment group showed higher consistency in relative scores of subjects than the placebo responder group. The ϕ for all 30 items averages 0.649, 0.656, and 0.667, respectively, also denoting moderate reliability across items.

DISCUSSION

This study uniquely applied the G theory to assess the extent to which each facet (ie, raters, subjects, and visits) contributed to the variability and inconsistency in subscale item scores of the PANSS for subjects who participated in a failed clinical trial. The study estimated how consistent ratings by different raters are, across visits when averaged across the items of the PANSS. Whereas a substantial finding is that of rater variability, the most salient factor contributing to error variance in the efficacy outcome, the G theory identified 3 other distinct substantial findings from our research using this data set.

First, the placebo response rate observed was 45.07% (32/71 subjects), and the treatment response rate was 60.28% (85/141 subjects). Placebo response is recognized to be markedly elevated in psychiatric clinical trials.[23] A review of placebo response of clinical trials in schizophrenia was performed by Kemp et al[24] and noted that the nature of the problem may be related to unreliable assessments conducted by site raters, as well as a range of other potential sources including poor trial design and poor adherence to medication. Kinon et al[23] report that among recent clinical trials in schizophrenia, the mean percentage of placebo responders was 25.0%, with a range of 0.00% to 41.0%. Whereas this was only a brief review, the

TABLE 1. Three-Facet Subject × Rater × Visit: Component Variance and Percentage of Variance for PANSS Positive Subscale Scores

	P1 Delusions	P2 Conceptual Disorganization	P3 Hallucinatory Behavior	P4 Excitement	P5 Grandiosity	P6 Suspiciousness/ Persecution	P7 Hostility
Placebo responders, source of variation							
Subject	0.029	0.031	0.030	0.109	0.000	0.079	0.199
	2%	3%	3%	3%	0%	6%	16%
Rater	0.119	0.116	0.190	0.214	0.099	0.021	0.014
	72%	69%	70%	72%	59%	50%	33%
Visit	0.102	0.011	0.030	0.002	0.095	0.099	0.096
	3%	4%	6%	3%	10%	11%	11%
Subject × rater	0.051	0.010	0.059	0.032	0.064	0.064	0.069
	8%	4%	10%	6%	11%	11%	14%
Subject × visit	0.000	0.035	0.021	0.025	0.029	0.035	0.039
	0%	8%	5%	5%	7%	8%	11%
Rater × visit	0.100	0.059	0.040	0.051	0.061	0.065	0.069
	15%	12%	6%	11%	13%	14%	15%
Ep^2	0.453	0.610	0.612	0.910	0.529	0.631	0.961
ϕ	0.451	0.559	0.542	0.689	0.501	0.542	0.873
Placebo nonresponders, source of variation							
Subject	0.030	0.035	0.033	0.112	0.010	0.081	0.186
	3%	4%	4%	5%	1%	7%	18%
Rater	0.120	0.119	0.201	0.200	0.099	0.024	0.024
	65%	64%	68%	65%	55%	48%	31%
Visit	0.105	0.100	0.046	0.102	0.102	0.101	0.099
	5%	8%	10%	9%	11%	13%	14%
Subject × rater	0.054	0.019	0.061	0.089	0.076	0.069	0.039
	10%	5%	10%	8%	13%	12%	15%
Subject × visit	0.000	0.045	0.035	0.026	0.060	0.043	0.042
	1%	7%	4%	4%	5%	6%	9%
Rater × visit	0.109	0.067	0.042	0.054	0.075	0.071	0.059
	16%	12%	4%	9%	15%	14%	13%
Ep^2	0.479	0.612	0.633	0.846	0.504	0.623	0.935
ϕ	0.462	0.605	0.601	0.687	0.501	0.551	0.901
Treatment group, source of variation							
Subject	0.050	0.046	0.040	0.136	0.011	0.097	0.187
	6%	6%	7%	7%	3%	11%	18%
Rater	00.130	0.136	0.199	0.210	0.097	0.036	0.035
	60%	59%	60%	59%	49%	44%	29%
Visit	0.099	0.114	0.069	0.112	0.103	0.112	0.100
	6%	10%	12%	10%	18%	15%	16%
Subject × rater	0.056	0.026	0.064	0.098	0.102	0.079	0.038
	10%	7%	7%	6%	13%	12%	16%
Subject × visit	0.060	0.051	0.053	0.035	0.069	0.054	0.051
	4%	9%	9%	8%	8%	9%	10%
Rater × visit	0.106	0.079	0.049	0.067	0.073	0.082	0.062
	14%	9%	5%	10%	9%	9%	11%
Ep2	0.500	0.600	0.612	0.900	0.596	0.652	0.900
ϕ	0.496	0.611	0.611	0.856	0.512	0.543	0.899

Ep^2 indicates generalizability coefficient; ϕ, index of dependability.

placebo response rates for the present study are greater than the range in the review by Kinon et al.[23] Moreover, Quitkin et al[25] stated that sudden fleeting responses observed during pharmacological treatment are in all likelihood the result of placebo effects. Although some response in placebo-treated patients might be expected in a clinical trial owing to the natural course of the illness, it is important to be able to identify when an uncharacteristic placebo response could influence the results. Defining the root causes of failure of treatment interventions is a necessary first step in an attempt to remediate the problem.

Second, for a scientific investigation of trial failure to advance, it is important to establish whether a placebo response in

TABLE 2. Three-Facet Subject × Rater × Visit: Component Variance and Percentage of Variance for PANSS Negative Subscale Scores

	N1 Blunted Affect	N2 Emotional Withdrawal	N3 Poor Rapport	N4 Passive Apathetic Social Withdrawal	N5 Difficulty in Abstract Thinking	N6 Lack of Spontaneity and Flow of Conversation	N7 Stereotyped Thinking
Placebo responders, source of variation							
Subject	0.156	0.097	0.100	0.009	0.187	0.198	0.185
	7%	4%	5%	2%	10%	14%	10%
Rater	0.120	0.109	0.118	0.097	0.123	0.090	0.106
	69%	61%	65%	58%	69%	54%	60%
Visit	0.041	0.062	0.002	0.097	0.098	0.074	0.088
	2%	5%	1%	8%	10%	6%	7%
Subject × rater	0.060	0.054	0.089	0.081	0.009	0.067	0.071
	6%	5%	12%	12%	4%	8%	11%
Subject × visit	0.052	0.087	0.000	0.081	0.065	0.076	0.071
	4%	10%	0%	7%	5%	6%	6%
Rater × visit	0.067	0.091	0.099	0.061	0.002	0.061	0.009
	12%	15%	17%	13%	2%	12%	6%
Ep^2	0.911	0.452	0.497	0.421	0.941	0.496	0.813
ϕ	0.803	0.449	0.491	0.413	0.879	0.419	0.742
Placebo nonresponders, source of variation							
Subject	0.164	0.101	0.102	0.010	0.199	0.201	0.197
	9%	8%	7%	9%	15%	13%	12%
Rater	0.134	0.116	0.119	0.099	0.169	0.101	0.112
	64%	59%	60%	50%	59%	49%	54%
Visit	0.059	0.087	0.024	0.101	0.101	0.100	0.097
	5%	7%	3%	11%	8%	10%	10%
Subject × rater	0.068	0.064	0.097	0.100	0.012	0.068	0.075
	8%	7%	12%	11%	6%	10%	10%
Subject × visit	0.062	0.091	0.010	0.098	0.087	0.097	0.089
	5%	9%	4%	10%	6%	10%	7%
Rater × visit	0.079	0.099	0.100	0.079	0.046	0.076	0.012
	9%	10%	14%	9%	6%	8%	7%
Ep2	0.912	0.512	0.500	0.511	0.915	0.501	0.815
ϕ	0.816	0.511	0.486	0.468	0.901	0.459	0.785
Treatment group, source of variation							
Subject	0.168	0.106	0.113	0.014	0.201	0.212	0.194
	10%	9%	9%	12%	17%	15%	14%
Rater	0.138	0.118	0.124	0.102	0.178	0.106	0.113
	56%	55%	55%	48%	55%	45%	52%
Visit	0.066	0.092	0.025	0.109	0.115	0.101	0.101
	8%	8%	5%	10%	9%	11%	9%
Subject × rater	0.069	0.068	0.098	0.112	0.019	0.064	0.086
	10%	9%	8%	10%	7%	10%	9%
Subject × visit	0.071	0.094	0.101	0.099	0.089	0.092	0.099
	8%	10%	8%	10%	6%	11%	8%
Rater × visit	0.080	0.101	0.102	0.087	0.059	0.073	0.026
	8%	9%	13%	10%	6%	8%	8%
Ep2	0.900	0.513	0.511	0.516	0.909	0.526	0.856
ϕ	0.811	0.506	0.467	0.497	0.897	0.469	0.795

psychopathologic ratings is reliable or if there is a response to a single placebo administration. There may also be a placebo response to the repeated administration of a similar placebo in similar conditions,[26] for example, inconsistencies in ratings, symptoms, or visits. Correll et al[27] observed that early nonresponse to treatment, as measured by a 20% reduction in a Brief Psychiatric Rating Scale (BPRS) total score at week 1, predicted nonresponse at 4 weeks for 100% of patients. Other studies have suggested that early nonresponse to treatment within the first 2 weeks of treatment initiation is a good indicator of treatment refractoriness.[27–29] However, these studies failed to consider the source for placebo response. Moreover, early improvements to placebo treatment in psychiatric clinical trials have also been related to sustained clinical response at

TABLE 3. Three-Facet Subject × Rater × Visit: Component Variance and Percentage of Variance for PANSS General Psychopathology Subscale Scores

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16
Source of variation placebo responders																
Subject	0.101	0.086	0.051	0.084	0.068	0.068	0.064	0.051	0.004	0.054	0.065	0.077	0.043	0.008	0.075	0.066
	21%	13%	8%	13%	10%	10%	10%	8%	1%	6%	10%	12%	5%	2%	12%	10%
Rater	0.12	0.129	0.146	0.13	0.181	0.199	0.148	0.138	0.119	0.128	0.121	0.141	0.133	0.144	0.122	0.124
	33%	49%	61%	52%	66%	70%	62%	58%	31%	50%	43%	58%	56%	59%	34%	34%
Visit	0.006	0.003	0.001	0.004	0.002	0.003	0.005	0.005	0.017	0.008	0.005	0.006	0.001	0.005	0.003	0.002
	5%	3%	1%	3%	2%	2%	5%	4%	16%	8%	3%	3%	1%	3%	2%	2%
Subject × Rater	0.01	0.004	0.011	0.007	0.006	0.003	0.008	0.007	0.006	0.004	0.009	0.008	0.008	0.003	0.002	0.002
	8%	4%	8%	5%	4%	3%	6%	5%	4%	3%	8%	7%	7%	3%	2%	1%
Subject × visit	0.012	0.006	0	0.003	0.008	0.003	0.003	0.007	0.008	0.007	0.015	0.002	0.004	0.006	0.001	0.124
	9%	4%	0%	3%	6%	2%	2%	5%	6%	5%	10%	2%	3%	4%	1%	43%
Rater × visit	0.118	0.121	0.115	0.117	0.009	0.101	0.102	0.112	0.13	0.124	0.123	0.11	0.126	0.128	0.134	0.008
	24%	27%	22%	24%	12%	13%	15%	20%	42%	28%	26%	18%	28%	29%	49%	10%
Ep^2	0.512	0.511	0.802	0.498	0.846	0.901	0.891	0.888	0.512	0.965	0.746	0.689	0.772	0.587	0.539	0.719
ϕ	0.496	0.499	0.746	0.452	0.765	0.884	0.807	0.673	0.501	0.899	0.659	0.603	0.624	0.581	0.521	0.633
Source of variation Placebo nonresponders																
Subject	0.124	0.144	0.074	0.107	0.111	0.089	0.069	0.089	0.058	0.059	0.076	0.082	0.059	0.009	0.101	0.102
	20%	15%	10%	15%	13%	11%	12%	12%	5%	7%	13%	15%	10%	6%	15%	13%
Rater	0.122	0.136	0.154	0.132	0.175	0.187	0.156	0.149	0.126	0.138	0.127	0.146	0.138	0.145	0.136	0.128
	22%	43%	51%	43%	54%	62%	55%	50%	18%	49%	39%	52%	49%	52%	33%	28%
Visit	0.009	0.004	0.005	0.006	0.006	0.004	0.007	0.006	0.019	0.009	0.008	0.005	0.009	0.006	0.007	0.006
	12%	6%	8%	8%	7%	6%	8%	8%	13%	9%	6%	8%	1%	7%	4%	2%
Subject × rater	0.012	0.009	0.016	0.009	0.008	0.007	0.009	0.006	0.007	0.008	0.101	0.009	0.008	0.006	0.007	0.005
	12%	6%	9%	8%	8%	6%	8%	7%	13%	4%	10%	6%	13%	5%	2%	1%
Subject × visit	0.013	0.008	0.001	0	0.008	0.005	0.006	0.007	0.005	0.005	0.019	0.004	0.007	0.006	0.003	0.129
	12%	7%	6%	6%	8%	5%	7%	8%	12%	5%	10%	4%	3%	6%	3%	45%
Rater × visit	0.128	0.123	0.119	0.201	0.102	0.105	0.105	0.116	0.145	0.135	0.136	0.115	0.129	0.138	0.157	0.009
	22%	23%	16%	20%	10%	10%	10%	v15%	39%	26%	22%	15%	24%	24%	43%	11%
Ep2	0.523	0.516	0.9	0.521	0.846	0.906	0.902	0.905	0.536	0.967	0.815	0.756	0.773	0.602	0.564	0.812
ϕ	0.521	0.501	0.802	0.512	0.801	0.894	0.812	0.715	0.524	0.845	0.705	0.647	0.638	0.602	0.536	0.645
Source of variation treatment group																
Subject	0.127	0.147	0.078	0.12	0.124	0.094	0.072	0.087	0.054	0.061	0.075	0.079	0.061	0.01	0.099	0.108
	19%	18%	15%	15%	15%	13%	14%	14%	7%	8%	16%	15%	12%	8%	14%	13%
Rater	0.129	0.154	0.139	0.168	0.187	0.183	0.167	0.153	0.135	0.147	0.135	0.149	0.157	0.164	0.169	0.175
	19%	40%	41%	40%	49%	54%	50%	46%	20%	43%	37%	49%	46%	49%	31%	25%

Visit	0.01	0.008	0.006	0.007	0.002	0.003	0.007	0.007	0.02	0.012	0.014	0.008	0.016	0.003	0.008	0.017
	14%	7%	10%	9%	9%	8%	10%	10%	14%	10%	6%	8%	5%	7%	5%	1%
Subject × rater	0.016	0.099	0.016	0.046	0.007	0.009	0.101	0.005	0.008	0.006	0.099	0.009	0.008	0.007	0.008	0.003
	13%	6%	10%	8%	9%	8%	9%	7%	13%	8%	11%	9%	13%	7%	3%	2%
Subject × visit	0.018	0.009	0.008	0.001	0.005	0.006	0.008	0.009	0.003	0.004	0.02	0.008	0.006	0.007	0.006	0.131
	15%	7%	10%	10%	9%	7%	7%	10%	14%	6%	12%	4%	3%	8%	5%	43%
Rater × visit	0.135	0.134	0.12	0.199	0.109	0.11	0.106	0.113	0.156	0.134	0.139	0.136	0.135	0.141	0.161	0.101
	20%	22%	14%	18%	9%	10%	10%	13%	32%	25%	18%	15%	21%	21%	42%	16%
Ep2	0.546	0.513	0.902	0.536	0.905	0.912	0.908	0.916	0.601	0.946	0.864	0.805	0.802	0.711	0.569	0.873
Φ	0.531	0.469	0.872	0.516	0.812	0.864	0.862	0.835	0.536	0.856	0.735	0.712	0.726	0.608	0.496	0.621

study end point,[30] whereas our data show a graduated increase until visit 5 or visit 6. Pattern analysis of an initial clinical response to placebo has been used to differentiate between pharmacological drug treatment and placebo responders at study end points.[31,32] The errors in ratings are only one aspect of trial failure; other possible factors are implicated for failure of antipsychotic trials.

Third, we provide a practical example of how the G theory can be used to estimate the various sources of measurement error in psychopathology assessment using the PANSS. Our results demonstrates how these variances can then be used to enhance a data monitoring mechanism that minimizes error for a particular purpose, in this case, rater scores or ratings required for a generalizable result. Although the percent variability over time was good for most PANSS items, our findings indicated that for all PANSS positive items, the most variability was observed for raters (33%–72% for placebo responders, 21%–68% for placebo nonresponders, and 29%–60% for the treatment group), followed by variability of the interaction between rater and time point where 5 items (P1; P2; P4; P5, grandiosity; P6) showed variability ranging from 5% to 16%. Similarly, all items of the negative subscale showed the most percent variability for raters, followed by 4 items (N1, N2, N3, and N4) with percent variability for interaction between raters and visit. Fifteen of the 16 items of the general psychopathology subscale also showed the highest percent variability for raters. These results indicate that the source of unreliability are primarily found with raters, followed by the interaction between raters and visits, across all treatment groups with higher error variances noted for the placebo responder groups, suggesting that rater variability in scoring the PANSS is key to obtaining reliable efficacy measures.

Most typically, measurement has been considered from the perspective of classical test theory (using interrater reliability and internal consistency). Although these methods are useful in identifying the degree of precision with which we are able to administer and score the PANSS, the information provided by classical test theory estimates does not allow researchers and clinicians to propose ways in which errors may be reduced and PANSS scoring improved. One of the benefits of the G theory used in this study is that it guides the researcher and clinician regarding how to improve an efficacy measure through identification of error sources, rather than simply indicating overall weakness. A primary requirement of response to data monitoring intervention using the G theory is that problem sources of ratings on the PANSS (eg, raters, number of visits, and subjects) can be effectively and proactively identified and then assessed throughout the trial to determine whether intervention efforts are successful or alternative approaches are necessary. Therefore, using the G theory during the course of a clinical trial can assist ongoing data monitoring by examining the magnitude of variance components at precise visits throughout a study and take action to control for the raters' effect if variance components are noted (eg, through retraining). For example, items which show high rater variability (≥50%), for example, P1; P2; P3; P4; N1; N2; N3; N5; N7; G3, guilt feelings; G5, manneriisms and posturing; G6; and G7, motor retardation, would be expected to warrant additional review of scoring, administration, in-study data monitoring, and supplemental training throughout the study. Some items are consistent with findings of Santor et al[8] whose item response analysis of the PANSS found N5, N7, and G5 to function poorly.

For a clinician who wants to evaluate a patient's progress, the index of dependability can be informative, especially the smallest detectable difference. From the smallest detectable

difference, a clinician will know what differences need to be measured to conclude that real change has occurred rather than measurement error.

Our study presents several methodological limitations. First, a possible limitation may be sample bias, as we selected the placebo responders, the placebo nonresponders, and the treatment group from one study sample. This selection bias may limit the generalization of the conclusions of this study; however, the PANSS scores for this study are across the psychopathology spectrum, with a PANSS total score ranging from 70 or higher to 140 or lower. Similar samples of PANSS ratings from larger multicenter trials with different study drugs would be recommended. Second, the use of the G theory for estimating reliability of observational data has been disputed because of theoretical and practical problems. One of the main areas of interest is that observational data are usually gathered over the dimension of time. It should be noted that the data presented in this study was obtained 1 week apart and would not likely be affected by changes in true score, but data obtained from occasions that were more than 1 week apart might confound random error with changes in true scores.[11] A third limitation lies in the method of the G theory. Despite its advantages, because of the sampling variability inherent in sample data, the estimation method may give estimates for variance components that are negative, which is theoretically impossible. The likelihood of this happening increases when the design is unbalanced, and there are small frequencies in some cells, which was the case in the present study. There were instances when the estimated variance component was negative. When this occurred, it was set equal to zero, which is the recommended approach.[12] However, the few occasions of negative variances resulted in very small absolute values, which did not substantially affect the generalizations made. A fourth limitation lies in the assumptions of the G theory, which is that of stationarity, which is violated here because we know a priori that the scores are changing (eg, subjects are responders in both the placebo responder group and the treatment group); and thus, measurements over time are problematic in this study. The G theory offers a useful tool for operationalizing the efficiency of the PANSS. In particular, a researcher can select a suitable criterion for dependability a priori and then perform D studies to examine various combinations of error that can be formed to reach the criterion threshold. Creating several methods with acceptable levels of dependability has many benefits; researchers can quantify the cost of each source of error, raters can be retrained during the study if the source of error is identified as raters, and the results of intermittent G theory analysis can be used to inform selection. Further studies should be used to collect rating scale data across a greater number of time points to improve reliability of variance estimation. Finally, the lack of the demographic characteristics (level of training, professional degree, experience with rating the PANSS), although subject's age, ethnicity, and sex did not have an initial effect on the PANSS subscales and total scores.

Conclusions and Future Implications

The results presented here are indicative of a relationship between rater reliability and subsequent response the PANSS efficacy measure using clinical trial data. Concerns over the reliability of change scores have been the subject of discussion in psychology,[33] and several researchers have raised important questions about the applicability of reliability estimation to the study of change.[34] Our results confirm the efficacy of applying the G theory to the estimation of reliability of change over time as a supplement to other reliability approaches.

Numerous randomized clinical trials in schizophrenia have failed to separate active pharmacological treatment from placebo. A number of reasons have been presented, including rater training, interaction between subjects and raters, interaction between the study sponsor and sites, trial design, and poorly designed measurement instruments. Key among the reasons cited for failure is the adequacy of rater training. Minimizing the effect of placebo response in clinical trials compares to increasing the signal-to-noise ratio in the data.[23] Following principles of good research design (use of blinded raters, monitoring ratings throughout a study, and retraining raters during longitudinal studies) can make differences in active treatment and placebo more prevalent. The significance of rater training for reliability and validity in clinical trials is recognized as vital.[35] Training conducted at initial investigator's meetings may not be effective enough for longitudinal clinical trials and assessment of the level of psychopathology over time. Findings can be used to guide data monitoring, rater training, and identification of PANSS items, which may require supplementary training.

AUTHOR DISCLOSURE INFORMATION

WCY, GD, MO, AD, BR, SJ, and AK are consultants for, have received honoraria from, or have conducted clinical research supported by one or more of the following: National Institute of Mental Health, Hoffman LaRoche Inc, Otsuka Pharmaceuticals, Johnson & Johnson Pharmaceuticals, F Hoffman LaRoche Ltd, and Janssen Pharmaceuticals. In the past 5 years, no author has received reimbursements, fees, funding, or salary from an organization that may in any way gain or lose financially from the publication of this manuscript, either now or in the future. No authors hold any stocks or shares in an organization that may in any way gain or lose financially from the publication of this manuscript, either now or in the future. No authors hold or are currently applying for any patents relating to the content of the manuscript. No authors have received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript. All other authors have no competing funding interests. All authors have no nonfinancial competing interests (political, personal, religious, ideological, academic, intellectual, commercial, or other) to declare in relation to this manuscript.

REFERENCES

1. Fayers PM, Ashby D, Parmar MK. Tutorial in biostatistics Bayesian data monitoring in clinical trials. *Stat Methods*. 1997;16:1413–1430.
2. Jennison C, Turnbul BW. Statistical approaches to interim monitoring of medical trials: a review and commentary. *Stat Science*. 1990;5:299–317.
3. Fava M, Evins AE, Dorer DJ, et al. The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychother Psychosom*. 2003;72: 115–127.
4. Otto MW, Nierenberg AA. Assay sensitivity, failed clinical trials and the conduct of science. *Psychother Psychosom*. 2002;71:241–243.
5. Schatzberg AF, Kraemer HC. Use of placebo control groups in evaluating efficacy of treatment of unipolar major depression. *Biol Psychiatry*. 2000;47(8):736–744.
6. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull*. 1987;13(2):261–276.
7. Kay SR, Sevy S. Pyramidal model of schizophrenia. *Schizophr Bull*. 1990;16:537–545.
8. Santor DA, Ascher-Svanum H, Lindenmayer JP, et al. Item response analysis of the Positive and Negative Syndrome Scale. *BMC Psychiatry*. 2007;7:66–73.

9. Levine SZ, Rabinowitz J, Rizopoulos D. Recommendations to improve the Positive and Negative Syndrome Scale (PANSS) based on item response theory. *Psychiatry Res*. 2011;188(3):446–452.
10. Cone JD. The behavioral assessment grid: a conceptual framework and taxonomy. *Behav Ther*. 1978;9:882–888.
11. Brennan RL. *Generalizability Theory*. New York, NY: Springer-Verlag; 2001.
12. Cronbach LJ, Gleser CG, Rajaratnam N, et al. *The Dependability of Behavioral Measurements*. New York, NY: Wiley; 1972.
13. Shavelson RJ, Webb NM. *Generalizability Theory: A Primer*. Newbury Park, CA: Sage; 1991.
14. Nekolaichuk CL, Maguire TO, Suarez-Almazor M, et al. Assessing the reliability of patient, nurse, and family caregiver symptom ratings in hospitalized advanced cancer patients. *J Clin Oncol*. 1999;17(11): 3621–3630.
15. Bergeron R, Floyd RG, McCormack AC, et al. The generalizability of externalizing behavior composites and subscale scores across time, rater, and instrument. *School Psychology Review*. 2008;37:91–108.
16. Brennan RL. Performance assessments from the perspective of generalizability theory. *App Psych Measurement*. 2000;24:339–353.
17. Guy W. *ECDEU Assessment Manual for Psychopharmacology*. Rockville, MD: US Department of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration; 1976.
18. Leucht S, Kane JM, Kissling W, et al. Clinical implications of Brief Psychiatric Rating Scale scores. *Br J Psychiatry*. 2005;187:366–371.
19. Ascher-Svanum H, Nyhuis AW, Faries DE, et al. Clinical, functional, and economic ramifications of early nonresponse to antipsychotics in the naturalistic treatment of schizophrenia. *Schizophr Bull*. 2008:34: 1163–1171.
20. Suen HK, Lei PW. Classical versus generalizability theory of measurement. *Educational Measurement*. 2007:4;1–13.
21. Opler LA, Kay SR, Lindenmayer JP, et al. *SCI- PANSS*. Toronto, Canada: Multi-Health Systems Inc; 1992.
22. Andresen EM, Catlin TK, Wyrwich KW, et al. Retest reliability of surveillance questions on health related quality of life. *J Epidemiol Community Health*. 2003;57:339–343.
23. Kinon BJ, Potts AJ, Watson SB. Placebo response in clinical trials with schizophrenia patients. *Curr Opin Psychiatry*. 2011;24(2):107–113.
24. Kemp AS, Schooler NR, Kalali AH, et al. What is causing the reduced drug-placebo difference in recent schizophrenia trials and what can be done about it? *Schizophr Bull*. 2008;38:504–509.
25. Quitkin FM, Stewart JW, McGrath PJ. Further evidence that a placebo response to antidepressants can be identified. *Am J Psychiatry*. 1993; 150:566–570.
26. Kaptchuk TJ, Kelley JM, Conboy LA, et al. Components of placebo effect: randomised controlled trial in patients with irritable bowel syndrome. *BMJ*. 2008;336(7651):999–1003.
27. Correll CU, Malhotra AK, Kaushik S, et al. Early prediction of antipsychotic response in schizophrenia. *Am J Psychiatry*. 2003;160: 2063–2065.
28. Leucht S, Heres S, Hamann J, et al. Methodological issues in current antipsychotic drug trials. *Schizophr Bull*. 2008;34:275–285.
29. Kinon BJ, Chen L, Ascher-Svanum H, et al. Predicting response to atypical antipsychotics based on early response in the treatment of schizophrenia. *Schizophr Res*. 2008;102:230–240.
30. Gomeni R, Merlo-Pich E. Bayesian modelling and ROC analysis to predict placebo responders using clinical score measured in the initial weeks of treatment in depression trials. *Br J Clin Pharmacol*. 2007; 63(5):595–613.
31. Gibbons RD, Clark DC, Kupfer DJ. Exactly what does the Hamilton Depression Rating Scale measure? *J Psychiatr Res*. 1993;27:259–273.
32. Rothschild R, Quitkin FM. Review of the use of pattern analysis to differentiate true drug and placebo responses. *Psychother Psychosom*. 1992;58:170–177.
33. Rogosa DR, Willett JB. Demonstrating the reliability of the difference score in the measurement of change. *Journal of Education Measurement*. 1983;20:335–343.
34. Collins LM. Is reliability obsolete? A commentary on "Are Simple Gain Scores Obsolete?" *App Psych Measurement*. 1996;20:289–292.
35. Muller MJ, Wetzel H. Improvement of inter-rater reliability of PANSS items and subscales by a standardized rater training. *Acta Psych Scand*. 1998;98(2):135–139.