

Psychometric Properties of Report Card Instruments

Introduction

The technical appendix summarizes information on scoring and psychometric properties of the seven instruments included in the MHSIP Mental Health Report Card. These instruments are:

- C The Abnormal Involuntary Movement Scale (Table 1)
- C The Symptom Distress Scale (Table 2)
- C The Rosenberg Self-Esteem Scale (Table 3)
- C The Clinician Alcohol and Drug Use Scales (Table 4)
- C The Medical Outcomes Study 36-Item Short-Form Health Survey (Table 5)
- C The Child and Adolescent Functional Assessment Scale (Table 6)
- C The Consumer Satisfaction Survey (Table 7)

Important psychometric properties for assuring that the instruments consistently measure the constructs that they were intended to measure are validity and reliability. The appendix includes such types of validity and reliability as:

- C Face validity, or validity by assumption, that implies the obvious relevance of an instrument to the measured construct:
- C Convergent validity that shows whether the instrument in validation correlates with already established instrument for measuring the same construct;
- C Construct validity that shows whether the instrument in validation correlates with the constructs that are assumed to be related to the construct measured by the instrument in validation;
- C Discriminant validity that shows whether the instrument discriminates between the subjects know to differ in relation to the construct measured by the instrument in validation; or that shows whether the instrument does not correlate with constructs different from the one measured by the instrument under validation;
- C Inter-rater reliability that shows the extent of agreement among the raters;
- C Test-retest reliability that shows the extent of agreement among the several tests;
- C Internal consistency reliability using such statistics as Cronbach's alpha and Spearman-Brown split half coefficient.

The Abnormal Involuntary Movement Scale

The abnormal Involuntary Movement Scale (AIMS; Guy 1976) is a clinical rating instrument for tardive dyskinesia (TD). It includes individual body area scales that measure the severity of movement impairment in seven body areas, a total scale that results from summing individual body area ratings, and an overall severity scale that is a rating of the severity of all abnormal movements as a whole. A 5-point scale, where 0 stands for Anone,¹ 1 for Aminimal,² 2 for Amild,³ 3 for Amoderate,⁴ and 4 for Asevere,⁵ is used for rating both on the individual body area and on the overall severity scales.

The AIMS is a global rating method. Gardos et al. (1977) asserted that global rating methods of TD have high face validity, because the rate directly assesses dyskinetic movements. On the other hand, the authors emphasized the importance to demonstrate the reliability of the instruments, like the AIMS, in order to ensure that they are used in the intended way. Accordingly, table 1¹ presents studies that has provided evidence of the AIMS reliability.

When interpreting data from the table, it is important to bear in mind that TD severity of the assessed persons (Bergen et al. 1984) as well as experience and personality features of the raters (Smith et al. 1979; Lane et al. 1985) might have influenced the reliability estimates. More severe TD manifests more fluctuations, and hence, might have reduced test-retest reliability coefficients to a larger extent in the studies that employed more sick persons (Bergen et al. 1984). Since the AIMS does not provide any specific definition of the scoring scale points, both inter-rater and test-retest reliabilities might have been affected by the raters-characteristics. The AIMS requires the raters to compare the observed movement to the average movement disturbance seen in persons with TD. Such relative judgments may vary among raters with different backgrounds and personalities and may be more difficult of less experienced raters. Studies by Smith et al. (1979), Bergen et al. (1988), and Lane et al. (1985) empirically documented the rater-caused differences in scorings on the AIMS.

When reading table 1, attention should be paid also to limitations of the statistics used to estimate reliabilities. The Pearson correlation coefficient tends to overestimate agreement by not taking into account a constant difference in the scorings of one rater relative to another (Lane et al.) and of the variation of the subjects= TD severity (Bergen et al. 1984; Lane et al. 1985). Both coefficients decrease when a study is restricted to subjects with similar severity.

¹Only correlation coefficients that were significant at least on the .05 level are included in tables 1 through 7.

Table 1

Study	Subjects	Raters	Procedure	Inter-item Correlations	Inter-rater reliability (IRR)	Test-retest reliability (TRR)	Rater bias	Patient fluctuations
Chien et al. 1977		Experienced researchers in TD			IRR with the AIMS is higher than with several other rating instruments for TD			
Smiths et al. 1979	<ul style="list-style-type: none"> ⊆ Consumers of psychiatric inpatient services, about 85% of whom were receiving neuroleptics; ⊆ Subjects=mean total AIMS score was 5.36 with standard deviation (SD) of 4.79 ⊆ (N=377) 	<ul style="list-style-type: none"> ⊆ Well-trained raters ⊆ (N=4) 	<ul style="list-style-type: none"> ⊆ Two raters independently scored each subject; ⊆ Among different rater teams sample sizes varied from 39 to 48 subjects; ⊆ One rater team tested 35 subjects twice with approx. Seven week period between the tests 	For each rated person, r's among mean scores on the individual body area, total, and overall severity scales ranged from .23 to .88	For the individual body area, total, and overall severity scales, IRRS measured by the Pearson r's averaged over six two-rater teams ranged from .66 to .87	For the individual body area, total, and overall severity scales, approx. seven week TRRs measured by the Pearson r's averaged over six raters ranged from .40 to .75	One rater scored significantly higher than at least two others on the four individual items and the total scale	

<p>Bergen et al. 1984 (part 1)</p>	<p>⊆ Persons with mild TD receiving outpatient services, maintained on stable doses of neuroleptics; ⊆ subjects=mean total AIMS score was 8.25 with SD of 1.71 ⊆ (N=4)</p>	<ul style="list-style-type: none"> • Psychiatrists experienced in the assessment of TD • N=3 	<ul style="list-style-type: none"> • 11 to 13 examinations of persons with TD conducted within a month were video-recorded • Each rater completed from one to three scorings of each examination 				<p>Rater-s variability in re-rating subjects accounted for from 62 to 81 percent of the total variance attributable to both subjects=fluctuations in TD and re-ratings</p>	<p>⊆ Subjects=fluctuations in TD accounted for from 19 to 38 percent of the total variance attributable to both subjects=fluctuations in TD and re-ratings; ⊆ Subjects=variance estimates were not found reliable</p>
------------------------------------	--	--	--	--	--	--	--	--

Study	Subjects	Raters	Procedure	Inter-item Correlations	Inter-rater reliability (IRR)	Test-retest	Rater bias	Patient fluctuations
<p>Bergen et al. 1984 (part 2)</p>	<p>⊆ persons with TD whose severity ranged from equivocal to moderate ⊆ (N=22)</p>	<ul style="list-style-type: none"> • N=4 	<p>⊆ Each rater scored video-recorded examinations of 22 subjects; ⊆ One rater re-rated 19 of the same tapes six weeks later</p>		<p>⊆ For the individual body area and total scales, IRRs measured by the Pearson r's ranged from .48 to .97 ⊆ (N=22)</p>	<p>For the individual body area and total scales, six week TRR for one of the raters measured by the Pearson r's ranged from .63 to .95</p>		

<p>Lane et al. 1985</p>	<p>◦ Consumers of outpatient services, maintained on stable doses of neuroleptics; ◦ Subjects= total AIMS score was minimum 3 and at least one subscore was minimum 2 ◦ (N=33)</p>	<p>◦ Psychiatrists who had extensive experience with the AIMS and worked together for three years (N=2); ◦ Psychiatrists with minimal prior exposure to persons with TD (N=2)</p>	<p>◦ Persons with TD were examined on a weekly basis during a ten month period; ◦ Each person was rated independently by the raters present at the given session; ◦ At least three raters were present each but one week</p>		<p>For the individual body area, total, and overall severity scales, IRR measured by the intraclass r's averaged among four raters ranged from .50 to .79, while IRR measured by the Pearson r's ranged from .46 to .81</p>		<p>◦ During the whole study period, the experienced rater dyad had higher IRR in more body areas than did the inexperienced dyad; ◦ Experienced dyad did not reveal any systematic change in IRR, while agreement scores for inexperienced dyad went up more frequently than down</p>	
-------------------------	--	--	--	--	---	--	--	--

Study	Subjects	Raters	Procedure	Inter-item Correlations	Inter-rater reliability (IRR)	Test-retest reliability (TRR)	Rater bias	Patient fluctuations
Bergen et al. 1988	<ul style="list-style-type: none"> ○ Persons with a rather variable global severity and distribution of symptoms ○ (N=30) 	<ul style="list-style-type: none"> ○ A psychologist and a nurse; ○ Both without prior experience with TD 	<ul style="list-style-type: none"> ○ Examinations of persons with TD were video-recorded; ○ Both raters made five independent ratings of each examination; ○ Eight examinations were rated wt each of the rating sessions which took place three times per week during six weeks 		<ul style="list-style-type: none"> ○ For individual body area and overall severity scales, proportions of the persons with TD, for whom the two raters gave the same score, averaged over five examinations ranged from .53 to .79; ○ IRR was maintained on the same level over the whole study period; ○ AIMS scores given by both raters steadily increased with re-rating (p<.001) 		<ul style="list-style-type: none"> ○ Both raters scored with greater sensitivity to abnormality than the author of the study, an experienced researcher in TD; ○ The psychologist rated lower than the nurse on the two individual items and the total scale (p<.001) 	<ul style="list-style-type: none"> • Some persons were harder to rate than others (p<.05)

