Music and Medicine
2(1) 23-28

DOI: 10.1177/1943862109348616
http://mmd.sagepub.com
SAGE

Interrater Agreement on the Nordoff-Robbins Evaluation Scale I: Client-Therapist Relationship in Musical Activity

John F. Mahoney, MM, MA, NRMT, AMT, MT-BC, LCAT[1]

Abstract

The purpose of this study was to determine the degree of interrater agreement on an evaluation scale designed by Paul Nordoff and Clive Robbins to track the behavioral responses of autistic children undergoing individual improvisational music therapy and later understood to have application for a wider range of populations. The Client-Therapist Relationship in Musical Activity scale was first published in Nordoff and Robbins's book *Creative Music Therapy* in 1977, along with another scale entitled Musical Communicativeness. Although the scales are widely used, trials involving interobserver agreement have not been undertaken (Wigram, Nygaard, Pedersen, & Bonde, 2002). This study examined the revised version of the first scale, comparing the variance between the ratings of 10 video excerpts of music therapy sessions obtained from a group of certified music therapists ($N = 34$) who have practiced professionally for at least 3 years working with children with developmental delays or autism. Of the 34 participants, 21 had received advanced training and certification in the Nordoff-Robbins (NR) approach to music therapy and 13 had not received this training. The results showed that at a significance level of $p < .05$, 78% of the entire group of participants obtained mean scores that were within 1 point of the total group mean, 82% of the group with NR training obtained mean scores that were within 1 point of the NR group mean, and 74% of the group without NR training obtained mean scores that were within 1 point of the music therapist group mean.

Keywords

assessment scales, reliability study, music therapy, Nordoff-Robbins

In the field of music therapy, there is an ongoing need to develop reliable and valid assessment and evaluation procedures that ensure high professional standards. Such procedures are hallmarks in related disciplines such as psychology, speech and language therapy, and neurology (Wigram, Nygaard, Pedersen, & Bonde, 2002). Assessment and evaluation are important aspects of any therapeutic discipline not only because they help define theories of practice but also because they provide systems of clinical accountability. Assessment also plays a crucial role within the treatment process itself—it guides treatment and provides the basis for evaluation.

This study is an investigation of an evaluation scale used in Nordoff-Robbins music therapy (NRMT). NRMT, developed between 1959 and 1976, is one of the oldest music therapy models in existence (Wigram et al., 2002). The model is practiced on seven continents and has a global influence in the field. In this approach, music is improvised by client(s) and therapist based on the belief that pathology can be both expressed and bypassed in music (Aigen, 1995). In NRMT, music is used *as* therapy, meaning that the created music is the primary means of motivating and effecting the client's therapeutic growth, providing both a stimulus and a response medium for the therapy process to take place (Bruscia, 1987). In this approach the client takes an active role in the music making, and there is an emphasis on the therapist using as wide a tonal and harmonic language as possible. The most effective music emerges from the ongoing clinical experience with the individual client rather than being prescribed and imported into the clinical process.

Several assessment and evaluation tools have been developed for use in NRMT. They are: indexing, the Tempo-Dynamic Schema, the Thirteen Categories of Response, and three evaluation scales. In the process called *indexing*, clinical sessions are video- or audiotaped and later meticulously analyzed for musical content. Client music expressions and therapist music interventions are noted and correlated to numbered points on the tape. Client expressions include such items as tonal vocalizations, the tempo and organization of rhythmic responses, and observations regarding the client's melodic and

[1] Temple University, Philadelphia, PA, USA

Corresponding Author:
John F. Mahoney, 300 High Point Drive, Apartment 609, Hartsdale, NY 10530
E-mail: john.mahoney@temple.edu

rhythmic facility with the expressive elements of music (Aigen, 2004). Gradually, an individual repertoire of music for therapy is developed (Bruscia, 1987).

The Tempo-Dynamic Schema associates emotional states to musical expressivity as manifested through tempo and dynamic range. An emotional state may be considered ''pathological'' if it is inflexible and musically meaningless or ''normal'' if it falls within the boundaries of common musical practice (Bruscia, 1987).

The Thirteen Categories of Response provides a tool for describing responses of individual children beating a drum with the therapist at the piano and for describing how the child reacts both personally and musically to different musical idioms, elements, and moods (Bruscia, 1987; Nordoff & Robbins, 1965).

Finally, Nordoff and Robbins developed three ''evaluation'' scales. These scales were initially adapted from scales designed to evaluate changes in the behavior of autistic children in the therapeutic day care milieu. Developed by Ruttenberg, Dratman, Frankno, and Wenar (1966), they were first published as ''An Instrument for Evaluating Autistic Children,'' and then revised and retitled, ''A Behavior Rating Instrument for Autistic Children—BRIAC.'' The Nordoff-Robbins (NR) scales evolved into instruments designed to assess responses within the specific clinical situation of improvisational music therapy. The scales are not restricted, however, to one specific patient population. Scale ratings are viewed as reflective of, and hence potentially diagnostic of, a variety of emotional and organic disturbances. These scales are: (a) Scale I: Child-Therapist(s) Relationship in Musical Activity, which identifies observable characteristic behaviors that help to define the developmental level of the client-therapist relationship; (b) Scale II: Musical Communicativeness, which identifies 10 levels of communicativeness that provide a means for charting the development of a child's ability to use music as a tool for communication; and (c) Scale III: Musical Response, which provides separate hierarchic taxonomies for rhythmic and melodic forms, differentiated by their complexity. The scale for rhythmic forms was derived principally from work with drums, and the melodic forms were evaluated primarily as they emerged in therapy as products of vocalizations or singing.

This study is specifically concerned with Scale I: The Client-Therapist Relationship in Musical Activity. It is the most widely used of the NR scales (Wigram et al., 2002) and receives the most extensive attention of all of the assessment and evaluation tools (with the possible exception of *indexing*) during training in the NR approach.

Scale I initially identified 20 discrete musical behaviors intended to assess a child's capacity for interpersonal relating through music. These 20 categories of behaviors were grouped into 10 hierarchically arranged levels of behaviors of a participatory nature with 10 corresponding levels of interactive behaviors described as resistive. The scale introduced the concept that behaviors on the part of the client that appear to be intended to resist participation with the therapist in coactive musical activity may be considered equally as relational as those that are intentionally participatory in nature—to resist relational musical participation implies that there are, in fact, relational forces in play.

In an early revision of the scale, Level 10 was deleted from the scale when it was determined that it had more application to group music therapy process than it did to individual music therapy process. This left 9 levels of possible response (18 items) remaining. In its current version, the first 3 levels of response have been collapsed into a single level, leaving the remaining 7 levels (14 items) of discrete relational activity.

Accepted scientific protocol requires a tool or system of measurement utlilized in clinical practice to undergo study to see that it gives consistent results when applied to similar situations or circumstances. A key to determining the practical usefulness of an assessment or evaluation tool is to determine the reliability of the device among music therapists working both within the same clinical model, (e.g., NRMT) and among music therapists from the general population. Although there is a wealth of case study material in the literature concerning music therapy with children and considerable literature suggesting the value of music therapy for child development (Wilson & Roehmann, 1987), there have been few controlled studies of NRMT with handicapped children. (Aldridge, Gustorff, & Neugebauer, 1995). And although widely used by clinicians (Wigram et al., 2002), there have been no published studies or references to the NR assessment scales.

The questions addressed in this study were: What is the degree of interrater agreement on Scale I: Client-Therapist Relationship in Musical Activity? Do raters with more extensive training (certification in the NR approach) have higher interrater reliability than those not trained in this approach when using this scale?

Method

Participants

Participants in the study were composed of a convenience sample of faculty, staff, and graduate students enrolled in the music therapy programs at Temple University and New York University, the Nordoff Robbins Centers in New York and London, and Senzoku Gakuen Music College in Kawasaki, Japan. To assure continuity in the protocol of data collection, the researcher in New York and Philadelphia, Helen Patey in London, and Clive Robbins in Kawasaki adhered to a prewritten script of instructions read aloud to the participants and maintained the same parameters of time allowance between excerpts for the raters at each location as they made their rating selection determinations.

Forty-three participants were recruited; however, 1 was eliminated because of failure to meet the criterion related to years of experience working with the specified population, and 8 were eliminated because they did not complete the rating forms in their entirety. The remaining sets of ratings numbered 34. The total number of ratings (samples) included in the data analysis was 340.

Table 1. Descriptions of Clients on Video Excerpts

Excerpt number	Gender	Age	Diagnosis
1	M	3	Pervasive developmental delay (PDD), hyperactive
2	M	4	PDD, visually impaired
3	M	5	Attention deficit disorder (ADD), speech delays
4	M	19	Developmental delay (DD), autism, severe intellectual impairment
5	M	3	DD, agenesis of the corpus callosum
6	F	24	DD, speech delays
7	M	12	PDD, autism
8	F	6	PDD
9	M	12	Autism, moderate intellectual impairment
10	M	3	Cerebral palsy, speech delays (nonverbal)

All participants in the study met the following criterion: professional training in music therapy with at least 3 years of work with children with developmental delays or autism. Participants included music therapists from four countries: United States, Korea, Japan, and England. One participant, though not originally from Australia, received NR training in that location. Of the entire group of participants, 24 were women and 10 were men. Breakdown by subgroup was as follows: 8 male and 13 female music therapists trained in the NR approach, and 11 female and 2 male music therapists not trained in the NR approach. Data regarding participants' specific age, race, or location of permanent residence were not collected.

To protect the human rights of the participants, the study was submitted and approved by the University Internal Review Board (IRB). Each participant received and signed a consent form (available from author upon request). Principles of confidentiality regarding the identity of the clients portrayed in the video excerpts, as well as the identities of the participants who viewed and rated the excerpts, were carefully maintained. Releases were signed and collected from parents or legal guardians of all clients shown in the video excerpts. Clients' names were not disclosed to the raters.

Design

The study was constructed to measure the degree of consistency between the scores resulting from raters' observations of 10 music therapy excerpts assembled by the researcher, with the total scores being considered the unit of analysis.

Materials

The researcher compiled a videotape or DVD of 10 excerpts of actual individual music therapy sessions. The use of a common therapist in all of the excerpts was intended to eliminate the variables introduced by the inclusion of multiple therapists. The excerpts are 2 to 4 min in length and document 10 clients working with the researcher. The brief excerpts present "snapshots" of clinical music interventions rather than the evolution of an individual client's growth over time as the tool was originally envisioned in the scale's application. However, assembling shorter excerpts provided an opportunity to obtain a greater range of assessments (ratings) relating to a larger variety of clinical interventions occurring with a greater number of clients. Interventions for inclusion on the videotape excerpt were chosen by the researcher to include the widest possible range among the various levels of interpersonal relationship described in the scale. Three excerpts were selected from intake sessions.

Table 1 reports gender, age, and diagnosis of the clients shown in the video excerpts, as well as whether the videotape excerpt occurred during an intake session or during an ongoing course of therapy.

Procedures

The participants and researcher coordinated meeting times and locations that were convenient to spend the approximately 85 min required to listen to the instructions read aloud, sign and gather IRB release forms, view each video excerpt, and devote 5 min after viewing each excerpt to rate it. Raters were each provided the written criteria accompanying the scale as published. Identical instructions, background information regarding the use and application of the scale, and brief, introductory information regarding the video excerpts were read aloud to all participants. Specific information regarding clients' ages or diagnosis was not provided to the raters so that they would be able to view the excerpts as objectively as possible. Participants in the study rated the excerpts either individually or in small groups.

Data Analysis

The collected data were entered the into SPSS Version 13.5. Average frequencies were calculated for participant ratings of each excerpt. Means and standard deviations were computed for the entire group ($N = 34$) and for each of the two subgroups of NR-trained therapists ($n = 21$) and non-NR-trained therapists ($n = 13$) for each excerpt. In analyzing the data obtained from the ratings, the question immediately arose regarding the appropriate method of statistical analysis with which to determine interrater agreement. In contrast to the design of typical psychometric scales in which selection categories are comparatively few but clearly differentiated from one another (i.e., introvert-extrovert, happy-sad), the present scale offers multiple selection categories with multiple, subtle gradations of

Table 2. Summary of Analyzed Data

Video tape excerpt number	Total group mean ($N = 34$)	MT group ($n = 13$) and NRMT group ($n = 21$) means		p (mean difference)	Standard deviation	p (Levine)	$\% \leq \pm 1$ from mean	$\% \geq \pm 1$ from mean	Total group $\% \leq \pm 1$ from mean
Excerpt 1	3.91	MT	4.54	.00	0.97	.02	77	23	88
		NRMT	3.53		0.57		100	0	
Excerpt 2	4.54	MT	4.12	.02	0.87	.01	77	23	86
		NRMT	4.80		0.50		95	5	
Excerpt 3	2.66	MT	2.53	.24	0.43	.53	100	0	97
		NRMT	2.74		0.53		95	5	
Excerpt 4	2.48	MT	2.65	.34	0.58	.13	100	0	88
		NRMT	2.37		0.97		76	24	
Excerpt 5	3.84	MT	3.98	.44	0.95	.81	77	23	79
		NRMT	3.75		0.74		81	19	
Excerpt 6	3.91	MT	3.98	.70	1.21	.10	46	54	66
		NRMT	3.86		0.74		86	14	
Excerpt 7	4.26	MT	4.45	.25	0.87	.38	77	23	82
		NRMT	4.14		0.68		86	14	
Excerpt 8	2.98	MT	2.57	.15	1.42	.57	47	53	49
		NRMT	3.23		1.17		52	48	
Excerpt 9	5.33	MT	5.79	.01	0.74	.95	77	23	79
		NRMT	5.04		0.71		81	19	
Excerpt 10	4.14	MT	4.24	.67	1.32	.18	62	38	64
		NRMT	4.08		0.91		67	33	

MT, music therapists; NRMT, Nordoff-Robbins music therapists.
Column 1 reports video tape excerpt number. Column 2 reports grand mean score including all participants from both groups. Column 3 reports subgroup mean scores. Column 4 reports significance level of the differences between subgroup mean scores. Column 5 reports dispersion of scores around the mean in terms of NR scale point units. Column 6 reports results of Levine's test for unequal variances, calculating the f ratio of the larger variance divided by the smaller variance. The test determines whether the two variances are significantly different from one another and whether there is a larger spread around the mean by one group or another. Column 7 reports percentages of mean scores by group within 1 NR scale point of the grand mean (all participants, not adjusted for group size). Column 8 reports percentages of the subgroup mean scores that were located more than 1 point away from the mean of that subgroup. Column 9 reports percentages of total group mean scores that were located within 1 NR scale point of the grand mean (all participants, not adjusted for group size).

interactions and hierarchical differences. Individual rater means that fell between $\pm .5$ points of the group (or subgroup) means were considered to fall within acceptable limits of agreement (Kappa ± 1 scale score unit). The Kappa statistic proposed by Cohen is used to quantify the agreement between observers independently classifying the same n units into the same k categories. The statistic corrects for the agreement expected to result from chance alone. It is also a measure that adjusts the observed proportion of agreement and ranges from pc/(1 – pc) to 1, where pc is the expected agreement that results from chance. In analyzing the ratings obtained from this study, the researcher considered a score $\leq \pm 1$ scale rating to be within a statistically acceptable range of agreement.

Results

Adjusting for the difference in group size, 78% of the entire group of participants obtained average scores that were within 1 point of the total group mean, 74% of the group of the non-NR-trained participants obtained average scores that were within 1 point of the music therapist group mean, and 82% of the NR-trained participants had average scores that were within 1 point of the NR certified group mean. This information is presented in Table 3. All scores were at a significance level of $p < .05$.

Again adjusting for the difference in group size, the average number of times an NR-trained participant obtained a score that was outside the accepted 1 scale score point spread around the mean was 1.81. The average number of times a non-NR-trained participant obtained a score that was outside the accepted 1 scale score point spread around the mean was 4.2. Table 2 presents a summary of the analyzed data.

Discussion

The primary research question for this study was to determine the degree of agreement between professional music therapists using the Client-Therapist Relationship in Musical Activity scale. In recruiting participants for the study, the researcher looked for a robust number of music therapists with 3 years of professional experience working with children challenged with developmental delays or autism without regard to their orientation of practice method.

Although the scale was initially used by its authors as a means of articulating the musical relationship between a client and therapist over a course of therapy or over the length of a complete music therapy session, the researcher decided for this study to select a greater number segments of shorter duration (2-4 min each) to facilitate a greater number of samples for comparison rather than to include fewer excerpts of longer

Table 3. Averages of Scores ($p < .05$)

Average of NR group ratings located within $\pm$ 1 scale point around the mean	Average of MT group ratings located within $\pm$ 1 scale point around the mean	Average of total group ratings located within $\pm$ 1 scale point around the mean
82%	74%	78%

NR, Nordoff-Robbins; MT, music therapists.

duration, which would have produced fewer samples. The researcher discussed this decision with the scale's surviving author, Clive Robbins, who suggested that although the methodology employed for the study might be viewed as a potential delimitation, it suggests that therapists might use the scale effectively to rate several segments from a single session or break each session into several, equal segments before rating it. The researcher's decision to include excerpts with slightly differing lengths was based on prioritizing musical values over strict temporal constraints as there was no way of establishing a uniformly clocked excerpt length without cutting off or extending musical phrases or ideas without arriving at musically contrived results.

While analyzing the data, differences between NR-trained and non-NR-trained participants became apparent. Determining the degree of difference between NR-trained and non-NR-trained therapists evolved as a secondary research question. It should be noted that *t* tests for Excerpts 1 and 2 have significant Levine test results, indicating that the *t*-test assumption of homogeneity of variance has been violated, invalidating the *t* tests for these two excerpts when comparing groups.

In a random sample of participants who do not have formal training in using the Client-Therapist Relationship in Musical Activity scale, or who are not familiar with the NR model itself, it might be expected that there would be raters whose individual perspectives do not fit with the philosophical foundations implicit in the music therapy model from which the scale evolved. In fact, there was considerable postrating discussion with more than one non-NR-trained participant who expressed strong disagreement with the very notion of regarding behaviors of a developmentally challenged child as possibly being "resistive" in nature or even with the idea that musical activity can be broken into discrete categories at all, as is explicitly done in the scale. These theoretical differences might account for the discrepancies from the group mean of the raters with orientations other than those espoused in the NR model. In other words, a great deal of the difference between the two groups seems to have more to do with differences in theoretical orientation than with the actual scale itself. Ultimately, when using a scale as a tool of measurement, one must adhere to its delineated parameters to maintain a rationale for its continued use. When using a scientific instrument of measurement, to impose one's personal theoretical bias over the intended use of the instrument is a disservice to the instrument as well as to the field. For an evaluation tool to be clinically useful or meaningful, it is imperative that it be utlilized in accordance with its intended design. In a study of this size, even a single "outlier" has a significant impact on the resulting data analysis. In the case of this study, a 10 percentage point increase in the degree of reliability would have been indicated had the researcher opted to eliminate a single set of responses most distant from the mean. However, the analysis of the data as presented in this study includes all collected responses and does not eliminate any outliers.

It should be reiterated that although the raters in the study were chosen as a convenience sample, the clinical excerpts presented to the participants were chosen with a very specific intent. It was the aim of the researcher to select excerpts with the widest range of relational musical behaviors over the full scope of the assessment scale being studied. There was greater discrepancy in the ratings for Excerpt 8 by both groups than for any other excerpts. This excerpt was chosen for inclusion by the researcher as an example of exceptionally resistive musical behavior. The concept of resistiveness as being relational is an interesting concept that requires some consideration to grasp. Forty-eight percent of NR-trained music therapists' and 53% of non-NR-trained music therapists' means fell outside the range of ± 1 scale score point spread from their respective group means when rating this excerpt. Perhaps the concept of "relational resistiveness" is a topic that warrants more attention during the NR training as well as further explanation within the field overall.

In summary, results of this study indicate that there is a healthy degree of interrater agreement among NR-trained clinicians using the Client-Therapist Relationship in Musical Activity scale. Although the degree of agreement is not as strong between raters without NR training, the overall degree of interrater agreement among all clinicians indicates that the scale can be useful for clinicians working from different orientations and theoretical perspectives.

Participants in the study indicated strong interest in additional training in the use of the NR assessment scales, and this is probably a sound idea if the scales are to be most effectively utilized in the field of music therapy. This could be accomplished by the inclusion of more hours of supervised applied usage during NR certification training, by offering CMTE (Continuing Music Therapy Education) courses at music therapy conferences, and by offering classes at NR centers for interested music therapists who are unable to participate in a complete course of NR certification training.

Revisions to the original version of the Client-Therapist Relationship in Musical Activity scale have gradually adjusted the balance between the inclusion of multiple descriptive categories of behaviors and fewer, broader categories of response. These revisions have brought the scale into greater congruity with the design of commonly utilized psychometric scales in other modalities. The trade-off of reducing the number of

individual levels of participation included in this scale has been that with each collapse of categories, although the potential for a higher degree of reliability rises, the descriptions of participation become broader, resulting in a reduction in the degree of specificity of client response.

Declaration of Conflicting Interests

The author declared that he had no conflicts of interests with respect to his authorship or the publication of this article.

Funding

The author declared that he received no financial support for his research and/or authorship of this article.

References

Aigen, K. (1995). Cognitive and affective processes in music therapy with individuals with developmental delays: A preliminary model for contemporary Nordoff-Robbins practice. *Music Therapy*, *13*, 13-46.

Aldridge, D., Gustorff, D., & Neugebauer, L. (1995). A preliminary study of creative music therapy in the treatment of children with developmental delay. *The Arts in Psychotherapy*, *22*, 189-205.

Bruscia, K. (1987). *Improvisational models of music therapy.* Springfield, IL: Thomas Book Publishers.

Nordoff, P., & Robbins, C. (1965). *Music therapy for handicapped children: Investigations and experiences*. New York: Rudolf Steiner.

Nordoff, P., & Robbins, C. (1977). *Creative music therapy*. New York. John Jay Company.

Ruttenberg, B., Dratman, M., Frankno, J., & Wenar, C. (1966). A behavior rating instrument for autistic children—BRIAC. *Journal of the Academy of Child Psychiatry*, *5*, 453-479.

Wigram, T., Nygaard, I., Pedersen, L., Bonde, O. (2002). *A comprehensive guide to music therapy: Theory, clinical practice, research, and training*. London: Jessica Kingsley.

Wilson, F., & Roehmann, F. (1987). *Music and child development*. St. Louis, MO: MMB Music.

Bio

John Mahoney serves as the Director of Clinical Services for Heartsong Inc. in Westchester County, NY, and is a doctoral candidate in the Music Therapy Department at Temple University.