Article

"Sounds good, but... what is it?" An introduction to outcome measurement from a music therapy perspective

Neta Spiro, Giorgos Tsiris & Charlotte Cripps

ABSTRACT

"Sounds good, but... what is it?" This is a common reaction to outcome measurement by music therapy practitioners and researchers who are less familiar with its meanings and practices. Given the prevailing evidence-based practice movement, outcome measurement does *'sound good'*. Some practitioners and researchers, however, have a limited or unclear understanding of what outcome measurement includes; particularly with respect to outcome measures and related terminology around their use. Responding to the *"what is it?"* question, this article provides an introduction to such terminology. It explores what outcome measures are and outlines characteristics related to their forms, uses and selection criteria. While pointing to some debates regarding outcome measurement, including its philosophical underpinnings, this introduction seeks to offer a useful platform for a critical and contextual understanding of the potential use of outcome measures in music therapy.

KEYWORDS

outcome measures, measurement, terminology, introduction, music therapy

Neta Spiro, PhD, is Research Fellow in Performance Science at the Royal College of Music. She is an Affiliated Researcher at the Centre for Music and Science, Faculty of Music, University of Cambridge and was previously Head of Research at Nordoff Robbins England and Wales.

Email: neta.spiro@rcm.ac.uk

Giorgos Tsiris, PhD, is Head of Research at Nordoff Robbins Scotland, Senior Lecturer in Music Therapy at Queen Margaret University in Edinburgh, and Knowledge Exchange Fellow at the Centre for the Arts as Wellbeing, University of Winchester.

Email: gtsiris@qmu.ac.uk

Charlotte Cripps holds an MA in Music as Development and was previously a research team member at Nordoff Robbins England and Wales, where she is currently studying music therapy.

Email: charlotte.cripps@nordoff-robbins.org.uk

Publication history: Submitted 9 February 2018; First published 27 June 2018.

SETTING THE CONTEXT: A STORY, OUR POSITION AND SOME DEBATES

The music therapy service at the Butterfly Care Home is on the verge of closing down after failing to demonstrate evidence of its impact on the residents. Bob, the music therapist, together with his line manager and the Head of Complementary Therapies – all of whom see music therapy as valuable but struggle to persuasively communicate its effectiveness to funders – are having a meeting with an external consultant to help them out.

With a background in research that prioritises measurement of psychometric properties, Liz – the consultant – is well-versed in the evidence-based world and the use of outcome measures. Despite her lack of knowledge of the music therapy field, of Bob's improvisational approach and of how music therapy works in this setting, Liz proposes the use of a particular outcome measure. Although not music therapy-specific, this validated measure is being used widely to test the effectiveness of psychological interventions in care homes.

The wide use of this measure – which is already 'out there' – seems appealing to Bob and his colleagues. Using this measure is likely to be a more persuasive way of *showing* that music therapy 'works', and they hope that funders will take its results more seriously than previous internal service evaluation feedback and vignettes.

In his mind, Bob already knows that music therapy works. This measure will simply be the tool to finally *prove* it. This is actually in contrast with Liz's view and the measure's aim: to test *whether* music therapy is effective or not. The discrepancy in their assumptions is left unspoken in the meeting; perhaps giving the illusion of mutual understanding. In any case, everyone is excited!

As they get nearer to the 'nuts and bolts' of how this measure will be used, some basic questions emerge. To their surprise, Bob and his colleagues start realising that they do not actually know what an outcome measure is and how outcome measurement works. After 45 minutes in this meeting, Bob takes a deep breath and asks: *"Sounds good, but… what is it?"*

This fictional story may resonate with situations that music therapists and researchers face. Despite the inclusion of assessment- and research-related modules in contemporary music therapy training programmes, training approaches and emphases vary dramatically around the world (Ridder & Tsiris 2015; Stegemann et al. 2016). Therefore, it cannot be taken for granted that qualification in music therapy prepares professionals for understanding outcome measures and terminology associated with outcome measurement. This lack of understanding is acknowledged to varying degrees and can be played out in cases such as the opening story. Some music therapists, like Bob, who have limited understanding of outcome measures but yet are interested in learning about them, may have the courage to ask questions and try to understand what an outcome measure is and how it works. Some, however, may hesitantly remain silent, while others may not see it as their role to question or even to be part of the decision-making process regarding outcome measurement. In any case, there are diverse views on whether the use of outcome measures in music therapy is appropriate at all.

Given the prevailing evidence-based practice movement (Wigram & Gold 2012), music therapists are likely to come across outcome measures in their workplaces. A basic understanding of outcome measurement is thus vital, and this is what this article seeks to offer; we focus on the *"what is it?"* question – that some music therapists, like Bob, would like to find out more about. As such, we offer an introduction to terminology around outcome measurement from a music therapy perspective by considering examples from the field. For some readers, this may be seen as covering 'old ground' – given the number of related publications, many of which are much more detailed (e.g. Lyons et al. 1997; Trauer 2010). For others, terminology may be unfamiliar and less straightforward. In either case, we hope that the music therapy frame of this article is of value to the music therapy profession and discipline. This framing can offer some insights and a bridge to wider professional and research questions in the field, including philosophical considerations that underpin outcome measurement and the debates around it in music therapy. Indeed, a basic understanding of outcome measurement is a necessary resource for critical awareness and constructive engagement in such debates.

To this end, and while this is not our primary focus, we firstly set a context by outlining our position and writing voice, and by laying out some debates around outcome measurement. Then, we focus on what outcome measures are and outline characteristics related to their forms, uses and selection criteria. In the discussion, we point towards some broader questions regarding outcome measurement in music therapy. By revisiting some epistemological and ontological

considerations, we consider some possibilities and risks that outcome measures may present for the music therapy profession and discipline.

Our position and writing voice

Through our experience with different organisations that provide music therapy in diverse settings, we have been at the meeting point of research and practice where discussions between music therapists, service-users and other stakeholders, such as funders, emerge. In these discussions and given our diverse backgrounds in music psychology, music therapy and ethnomusicology respectively, we have become increasingly aware of the varying understandings and levels of familiarity with outcome measures, be they our own or those of others. Our position is that constructive dialogues regarding outcome measurement require a basic shared understanding of related terminology and of how outcome measurement works. Equally, informed debates should be based on critical reflection and not on rigid views on how knowledge is generated and what knowledge matters (Tsiris, Spiro & Pavlicevic 2018). This balanced understanding also needs to consider the possibilities and limitations of each approach in relation to its area of investigation.

The terminology used in this article reflects the language met in outcome measurement literature generally and within music therapy. Such terminology is met in measurement-related jargon which is often associated with statistical concepts (e.g. statistical power, efficacy), and in relation to the underlying conceptualisation of music therapy practice. Given the introductory scope of the article, we explain this terminology by presenting practices, ideas and situations – like the opening story – that may be familiar to music therapists. This includes an intentional shift between jargon and more everyday language in different parts of the article. Also, we use terms such as “intervention” and “recipients of care” or “patients” which are commonly used in outcome measurement to describe therapeutic practices. These terms do not necessarily reflect our ways of understanding music therapy which welcomes sociocultural thinking. Such thinking, where terms such as “practice” and “participants” are more commonly used, brings to the fore a sensitivity to context and is associated with the emergence of community- and culture-oriented approaches to music therapy (e.g. Ansdell & Pavlicevic 2010; Pavlicevic & Ansdell 2004; Stige 2002; Stige et al. 2010; Wood 2016). In this article, however, we are keen to explore and communicate outcome measurement in its own terms of reference.[1]

Some debates around outcome measurement

Music therapy has the potential to bring change to people’s lives. This view seems to form the foundation of the music therapy profession and is shared among different music therapy models. Explorations of how and whether change occurs and the nature of this change, however, seem to vary in terms of focus and methodological approach. These variations relate to numerous factors including the philosophical underpinnings of different music therapy models (Bruscia 1987; Spiro, Tsiris & Pavlicevic 2014; Trondalen & Bonde 2012), as well as individual music therapists’ training and work experiences. Bruscia’s (1987) seminal book *Improvisational Models of Music Therapy* is one of the first attempts to outline the philosophical orientations of different music therapy models and their relationship not only to practice but also to assessment and evaluation. Bruscia highlights, for example, the nonreferential nature of music therapy improvisation in Creative Music Therapy (also known as Nordoff-Robbins music therapy). In this context, improvisation is regarded as intrinsically meaningful without depending upon other parameters for its interpretation. This theoretical assumption translates into music-centred practices. It also has explicit implications in terms of understanding therapeutic goals as contained within the musical goals and in terms of assessing such goals by treating people’s musical responses as the primary source of data.[2] And although different music therapy models evolve, expand and become multifaceted over time, some of their original orientations remain influential in their attitudes towards practice and assessment. In line with the theoretical underpinnings of these

[1] For a discussion regarding the use of outcome measures in music therapy research, see Spiro, Tsiris and Cripps (2018). For an overview of such measures, see the online resource *Outcome Measures in Music Therapy* (Cripps, Tsiris & Spiro 2016).

[2] In line with their theoretical orientation and assumptions, Nordoff and Robbins developed the Nordoff-Robbins Rating Scales. After their first publication (Nordoff & Robbins 1977), a revised version of these scales was published (Nordoff & Robbins 2007), while more recently there have been some studies exploring the use of these scales in contemporary practices (Mahoney 2010; Spiro et al. 2016).

models, some researchers – depending on their orientation – may not conceive music therapy practice as an intervention with clear-cut clinical outcomes. This is the case, for example, in some improvisational and ecological approaches to music therapy where means and ends are seen as integrated (e.g. Aigen 2005, 2007, 2008; Ansdell & DeNora 2016; Tsiris 2008; Wood 2016). Such perspectives lead to a different kind of understanding of 'outcomes' which does not always sit comfortably within the outcome measurement paradigm of the evidence-based practice movement.

In addition to the different music therapy models (including their philosophical underpinnings), outcome measurement is varied according to its different contexts of application. Variation can be in terms of reason for measurement, description of measure and measurement methods. In daily music therapy practice, for example, a common reason to assess outcome is to learn more about the client(s), what their needs are and to what extent these have been addressed (Lipe 2015; Waldon 2016; see also Garland, Kruse & Aarons 2003). In some research contexts, outcome assessment is commonly part of understanding the connection between an activity or intervention (e.g. music improvisation) and its result, consequence or impact.

Despite this variety of reasons, the origins and uses of outcome measures are often associated with naturalistic approaches to knowledge. Such approaches tend to uncover underlying patterns, associations of inputs-outputs and some kinds of causal relationship (Waldon 2016). These naturalistic approaches seem to be at odds with the constructivist or hermeneutic orientations of many music therapy models (such as psychodynamic and analytical music therapy; see Bruscia 1987; Wigram 1999) which currently prevail at least in Europe (De Backer & Sutton 2014; Ridder & Tsiris 2015; Stegemann et al. 2016). This potential mismatch between the underpinning orientations of outcome measurement and those of music therapy models has formed a basis for debates. Three common arguments that have been raised by music therapists and other professionals from a sociocultural perspective (e.g. Ansdell 2006; DeNora 2006; Procter 2011; Wood 2015) are the following:

- By focusing on certain aspects or 'ingredients' of music therapy work, outcome measurement compartmentalises practice and distances it from its context.
- There are concerns regarding the generalisation of results from an artificially controlled environment to a naturally messy reality. This generalisation reflects a dangerous leap from 'efficacy' to 'effectiveness'[3] and is connected to the perceived risk in assuming music therapy's effectiveness (or lack of) by not considering other variables (e.g. interventions that a client may receive alongside music therapy).
- Outcome measurement is predicated on a cause-and-effect view of music therapy and as such is perceived to be weak in assessing the multiplicity and variability of outcomes that are possible given the *emerging* nature of many music therapists' aims and work.

Some of these critiques – with particular reference to the experimental situations within which outcome measures might be used – are summed up by music therapist Gary Ansdell and music sociologist Tia DeNora:

> "We suggest that the very bright, hygienic light of the experimental situation (and the implicit ontology of music and of health/illness associated with this situation) is probably the wrong kind of light for seeing what it is that music does and what it is that music is. We believe a different, softer (dimmer!) form of light is needed in order to perceive the subtle things that music does, to see it in its natural workings and in ecologically valid circumstances. And that a slower form of dwelling with music in situ can help us to see the variegated processes by which music helps." (DeNora & Ansdell 2014: 4)

Despite these critiques, there are multiple reasons that motivate music therapists to focus on outcome measurement. In addition to those who advocate for outcome measurement from an epistemological viewpoint, some use outcome measures to gain multiple perspectives on their work and/or to communicate it in a language that seems to be valued more by the medical and scientific communities.

[3] Whereas 'effectiveness' refers to the degree of beneficial effect of an intervention under real-world settings, 'efficacy' intends to show that "treatment affects outcomes through a well-controlled, frequently laboratory-style experiment" (Wigram & Gold 2012: 168). Effectiveness is used throughout this article as a broader term, but outcome measures can often be used in both types of investigations. A useful distinction between efficacy trials (explanatory trials) and effectiveness trials (pragmatic trials) can be found in Gartlehner et al. (2006).

A recent example where some of the aforementioned debates have been played out is the publication of the TIME-A randomised clinical trial (Bieleninik et al. 2017) and the responses that it has generated from the academic community and the media, as well as professional bodies (e.g. American Music Therapy Association 2017; Gold & Bieleninik 2018; Turry 2018a, 2018b; Wilson 2017). Interestingly, these debates have been partly triggered by the attention that this trial gained not only due to it being the largest study of its kind and published in a high-profile journal, but also due to its outcomes, which do not support the use of improvisational music therapy for symptom reduction in children with autism spectrum disorder. Outcome measurement and especially the rationale behind the choice of a particular outcome measure are important ingredients in these debates. In response to Turry's (2018a) critique, the researchers stated:

> "[Turry's] points fall into two main categories: first, what is the most appropriate outcome for music therapy for children with autism spectrum disorder, and second, how can improvisational music therapy be standardised meaningfully. Both points are interconnected through process–outcome relations.
>
> Choosing an appropriate outcome is one of the hardest tasks in designing trials. Music therapy targets a variety of outcomes, which may differ across clients and may also change as the client and therapeutic process develop. This may be especially pertinent in autism spectrum disorder, which is a very heterogeneous disorder." (Gold & Bieleninik 2018: 90)

Our view is that such considerations and dialogues are essential in the field, and for promoting a meaningful and balanced relationship between research-based practice and practice-based research. To date, these dialogues seem to happen mainly in response to studies with 'negative' outcomes. The current article and other similar endeavours, such as academic publications (e.g. DeNora 2006) and conference presentations (e.g. Procter 2018), hopefully encourage a proactive and constructive engagement in such dialogues.

OUTCOME MEASURES AND THEIR USES

In addressing the initial *"what is it?"* question, one needs to recognise that there are many types of outcome measures, focusing on different presenting features, different settings and patient groups. Below we explore two forms of outcome measures: non-patient and patient- based measures. We then focus on different features of outcome measures and explore various considerations (including psychometric properties) that determine the selection and use of such measures.

As mentioned above, descriptions of outcome measurement abound in the literature. In brief, and with a focus on healthcare-related literature, an outcome measure is commonly understood as a tool developed to quantify or assess the effectiveness or impact of an intervention in terms of its capacity to have a specific, desired effect on presenting features or symptoms of patients.

The targeted presenting features or symptoms vary according to the patient group for which each outcome measure is designed. They might concern physical symptoms (e.g. pain, mobility, hormone levels), cognitive levels, mental health functioning or quality of life. Although in any given case there may be many simultaneous presenting features or symptoms, outcome measures are not intended to offer comprehensive measurements of everything. Individual outcome measures are used as indicators of change in certain presenting features and their findings may or may not be related to those of other measures which focus on other presenting features. Measures are often intended to be comparable across a group of patients or situations and often rely on numerical or categorical information such as frequency of certain types of behaviour (see, for example, the Music Therapy Diagnostic Assessment measure; Oldfield 2006).

Outcome – together with structure and process – is seen as a core component of healthcare provision. Donabedian, who is considered the founder of the study of quality in healthcare and medical outcomes research, emphasises the importance of "identifying key features of medical care that are associated with favourable outcomes, so that these features can be preserved despite the constraints imposed by an increasingly cost conscious healthcare environment" (Donabedian 1966, cited in Gilbody, House & Sheldon 2003: 9).

Indeed, the purpose of measuring outcomes of an intervention is, ideally, not only to establish what works but also to improve the quality of care (Gilbody, House & Sheldon 2003). The use of outcome measures can inform understanding of cost-effectiveness and decision-making in terms of funding for different interventions. Bolton and Breen (1999: 503) argue that "the ways in which patient outcomes are measured is a central issue in the

decision-making process of future treatment and health care regimens". The consistent use of the same outcome measure or the use of compatible measures, in particular, can enable policy-makers to compare the effects of different interventions across different patient groups (Jones, Edwards & Hounsome 2012).

Overall, there are six principal uses of outcome measures in medical practice: i) healthcare policy evaluation, ii) healthcare evaluation, iii) making individual clinical decisions in routine medical practice, iv) economic evaluation and resource allocation, v) clinical audit, and vi) healthcare needs assessment, which includes monitoring the health and assessing the needs of a population (Gilbody, House & Sheldon 2003). Furthermore, outcome measures are increasingly used as part of basic research, i.e. research that endeavours to understand basic mechanisms or functions which could be psychological, physical or neurological.

Similar uses of measures occur in music therapy practice and research within and beyond medical settings. As shown in a review of 26 music therapy-specific measures (Spiro, Tsiris & Cripps 2018), two main categories of function – in addition to assessment – are identified without being mutually exclusive: i) clinical work and treatment planning, and ii) screening and diagnostic assessment. Also, in some cases, the assessment elements of the measures are related to particular aspects of their application setting. The Music Therapy Special Education Assessment Tool (Langan 2009), for example, assesses the music-therapeutic process and progress in relation to special education settings and curricula.

Certain trends in terms of the focus of outcome measures have emerged over the years, and these trends are connected to changes in the international scene of healthcare and economics. During the 20th century in particular, many Western countries experienced a rapid rise in life expectancy, accompanied by increased incidences and duration of chronic illnesses. In this context, mortality rates are no longer sufficient measures of healthcare quality (Ebrahim 1995), and there has been a shift from focusing on length of life to quality of life (Ware 1995). This shift is reflected in the focus of research studies and respective outcome measures. For example, the Cochrane review on music therapy for people with dementia (Vink, Bruinsma & Scholten 2003) identified a number of studies focusing on music therapy's impact on patients' depression and emotional wellbeing, both of which are connected to people's quality of life. This shift of focus has occurred particularly in healthcare whereby measures of population mortality and morbidity are being replaced with patient-based values surrounding health (McDaniel & Bach 1995; McDowell & Newell 1996). The focus of each outcome measure can be taken as an indication of what is valued by the developers (and users) of such measures, or of what they think will be valued by those who read its results. Measures developed specifically for music therapy commonly focus on communication and/or interaction, cognitive, physical, social and emotional aspects, as well as musical skills and participation. The latter is one of the distinctive foci of music therapy-specific measures. Examples of musical aspects that are measured include: length of playing and rhythmic synchrony (Grant 1995), sonorous musical communication (Raglio, Traficante & Oasi 2006), independent playing, unusual interest in structure or shapes of instruments (Oldfield 2006), and qualities of musical participation and resistiveness (Nordoff & Robbins 1977).

FORMS OF OUTCOME MEASURES

There are two main forms of outcome measures: non-patient-based and patient-based outcome measures.[4] Non-patient-based outcome measures predominantly assess impairments of a patient, whereas patient-based measures tend to focus on the impact that an impairment or injury may have on patients' daily lives (Michener 2011). For instance, rather than evaluating patients' subjective reports on mobility issues or the personal impacts of decreased mobility, patients' functional abilities might be measured using a non-patient-based measure such as the Barthel Index (Collin et al. 1988). This measure has been used in music therapy research by, for example, Raglio et al. (2010) to observe how well a patient can carry out activities of daily living.

On the other hand, patient-based measures are distinguished mainly through the data collection method, since they directly look to the patient to provide data. Despite the enormous array of such measures, patient-based outcome measures can be described as

> "questionnaires or related forms of assessment that patients complete by themselves or, when necessary, others complete on their behalf, in

[4] Non-patient-based measures are also known as proxy, non-patient reported or clinician-rated outcome measures. Patient-based measures are also known as self-reported measures.

> order that evidence is obtained of their experiences and concerns in relation to health status, health-related quality of life (QoL) and the results of treatments received" (Fitzpatrick et al. 1998: 1).

The fact that 'patient-based' might refer either to measures rated by the patients themselves or to measures rated by a third-party informant, such as a caregiver or a clinician, arguably creates some ambiguity in the classification of such measures. In any case, patient-based outcomes are particularly relevant to interventions, such as music therapy, that involve participation and development of patient-therapist relationship, and for this reason we discuss them in greater detail below. Firstly, however, we report on non-patient-based outcome measures which were commonly used in healthcare before the relatively recent emphasis on those which are patient-reported.

Non-patient-based outcome measures

Non-patient-based measures often do not require direct input by the patient. This can be very useful in instances where patients are not in a position, or lack the capacity, to discuss their experiences (e.g. people with advanced dementia or severe autism spectrum disorder). In such cases, measures might rely on task completion or observational methods, either completed by a clinician or someone else close to the patient (e.g. family member). Among several measures reported in Cripps, Tsiris and Spiro (2016; see also Spiro, Tsiris & Cripps 2018), an example of a music therapy-specific outcome measure which is non-patient-based is the Music Therapy Checklist (Raglio, Traficante & Oasi 2007).

A wide range of non-patient-based measures are used particularly in the area of dementia: these include task-based activities that would screen for dementia, such as the 7 Minute Screen (Solomon & Pendlebury 1998), and observational measures to quantify aggression in behaviour, such as the Empirical Behavioral Pathology in Alzheimer's Disease (E-BEHAVE-AD) rating scale (Auer, Monteiro & Reisberg 1996). In the area of autism, the Emotion Recognition Test (ERT) involves task completion whereby the child with autism is asked to identify what emotions are represented by standardised photographs of facial expressions (Ryan & Charragain 2010). Another autism-related outcome measure is the Autism Social Skills Profile (ASSP); a measure based on child observation that identifies social reciprocity, social participation and detrimental social behaviours. In Schwartzberg and Silverman's study (2007) the ASSP was completed by parents to examine the effects of music-based social stories on their children's comprehension and generalisation of social skills. None of these measures, however, were developed specifically for music therapy.

Measures that detect physiological features of the patient can be used to indicate emotional changes. For instance, plasma cortisol in saliva is a biochemical marker for stress (Chu et al. 2013). Taken together, clinician-rated measures and patient-based measures can be mutually informative and work in conjunction with each other.

Patient-based outcome measures

Patient-based outcome measures are particularly important given that they consider patients' perspectives: they enable people who receive or take part in a healthcare intervention to communicate their experience. It is also within the interests and priorities of service providers to obtain feedback and information directly from the service-users or treatment recipients. This is evident in the emphasis of healthcare systems on service-user involvement and in the corresponding outcomes movement (Barr 1995) which emphasises the need for patient-based measures which correspond appropriately to the complex nature of practices, such as the arts therapies (Hackett 2016). This emphasis on measuring the impact of healthcare interventions from the patients' perspectives led, for example, to the introduction of Patient Reported Outcome Measures (PROMs) within the UK's National Health Service (The Chartered Society of Physiotherapy 2013).

Patient-based measures differ from those developed in many biomedical contexts in terms of what they seek to measure. Whilst biomedical measures tend to monitor physiological changes, patient-based measures ask patients to feed back on "unavoidably 'subjective phenomena' that cannot be objectively verified" (Albrecht 1994, cited in Gilbody, House & Sheldon 2003: 10), such as patients' own experiences of satisfaction, difficulty, distress, health improvement or symptom severity. A similar distinction is made between patient-based and clinical measures. The latter seems to be "narrowly focused", principally used by health professionals to "assess physiologic, other biomedical, or limited functional dimensions of health" (Barr 1995: 13). On the other hand, patient-based outcome measures seem to be more broadly defined and focus more on patients' values and

perceptions concerning their own health. Thus, patient-based measures often address aspects of health that are related to quality of life and health, including psychological, social and physical health, impairments, functional status, health perceptions and opportunities (Testa & Nackley 1994).

In sum, the patient-oriented focus of such measures characterises how data is collected, as well as what data is produced. Their data collection methods often include questionnaires, interview schedules and rating scales. The Hospice Music Therapy Assessment (Maue-Johnson & Tanguay 2006) is an example of a music therapy-specific outcome measure where data collection includes interviews with the patient and their family members.

FEATURES INFORMING CHOICE AND USE OF OUTCOME MEASURES

Having set the wider context of outcome measurement and presented the two main forms of measures, here we focus on the features that determine their use. In determining whether an outcome measure is appropriate and relevant for use within a given practice or research context, and informed by the work of Bausewein et al. (2011), we propose six key considerations:

(i) Aims of use: The aim of any assessment informs the duration of the enquiry, the type of data, as well as the expertise required to carry out the assessment. The chosen outcome measure must be suitable for the ultimate aims of an assessment.

(ii) Accessibility: This concerns the availability, cost, complexity, as well as length of time expected to get access to and administer a given outcome measure. Although, some measures may be open-access, many need to be purchased. Decisions regarding the pricing of outcome measures usually lie with the developers and their affiliated institutions. Whilst some outcome measures can be self-administered, some measures, such as the Music Therapy Assessment Tool for Awareness in Disorders of Consciousness (MATADOC; Magee 2007), require training to administer and are only available to the trained or initiated user, whether practitioner or researcher. Also, some measures, such as the Music Therapy Assessment for Disturbed Adolescents (Wells 1988), are task-based and require the administration of a specific protocol.

(iii) Categories of outcome: This refers to the specific kind of change that a measure aims to monitor. This might include, for example, levels of agitation, quality of life, or pain severity. What needs to be measured is informed by the purpose of each enquiry. In other words, outcome measures should be congruent with the reasons for using them. Along these lines, Bausewein et al. (2011) suggest that when selecting which measure to use one must consider what the measurement data would be used for. For instance, is the measure for a research study or for routine clinical purposes?

(iv) Type of assessment scale: The assessment scale type needs to be considered carefully, alongside factors pertaining to the use of such a scale in real-life contexts and with particular populations. For example, a highly sophisticated and complex measuring scale may not be appropriate for routine clinical checks administered by busy hospital staff. Likewise, a rating scale that requires clinicians to ask complex verbal information from a cognitively impaired patient would be problematic. In all cases, the viability of data collection methods should be ethically sound.[5]

(v) Condition group: Condition group concerns the classification of symptoms as they appear in different patient groups. Such classification influences decisions regarding what type of data might be desirable and what data would be realistic to be expected. Outcome measures are commonly developed for patient groups with specific symptoms or presenting features. For example, in the context of disorders of consciousness, the following aspects are commonly focused on: motor responses, arousal, as well as auditory and visual responsiveness. An example of a music therapy outcome measure assessing these aspects is the MATADOC (Magee 2007).

(vi) Disciplinary origin: The purpose and the approach behind the design of an outcome measure is typically influenced by its respective target field of practice. The scale Interest in Music (IiM; Gold et al. 2013), for example, was developed within the field of music therapy to measure interest in music among clients in mental health care. The purpose and approach of this scale have been influenced by contextual and relational music therapy models which propose the importance of music-related outcomes in clients' everyday lives.

In addition to the six key considerations mentioned above, the selection and use of appropriate outcome measures needs to be underpinned by a number of practical factors such as: the suitability of a measure for a given practice

[5] For a discussion of research ethics considerations in music therapy and in arts and health more broadly, see Farrant, Tsiris and Pavlicevic (2014).

or research situation, including its context and patient group (location, diagnoses, symptoms, age range, cognitive capabilities); purpose, methods of data collection, ease of administration, accessibility, cost, length, and interpretability, as well as internal consistency and a theoretical fit between what is being measured and the measuring instrument itself. In addition to these practical considerations, equally important in the selection and use of outcome measures are a range of conceptual and technical features regarding outcome measurement. These features – each of which are developed in the respective subsections below – relate to sample size, measurement of multifaceted phenomena, context specificity vs. comparability, as well as feasibility and psychometric properties of measures.

Sample size

The acceptability and, where relevant, the statistical power[6] of outcome measurement results is often associated with sample size (Guo, Chen & Luh 2011), and various music therapy studies have been criticised for their small sample size. Music therapy is not the only field in which questions around sample size, statistical methods and reporting have arisen (for examples in other fields see Button et. al. 2013; Ioannidis 2005). Though the criticism of many studies concerns small sample sizes, the assumption that larger samples lead automatically to stronger findings has been widely debated.[7] Given that, in many cases, outcome measures are used in very specific circumstances there is no necessary assumption that results from a given outcome measure are generalisable beyond their specific aspects; neither is there an inherent restriction on looking at individual differences in the context of outcome measures. Although sample size is a common research concern, related questions may arise in relation to the number of participants for whom outcome is measured in practice contexts. Similar questions around sample size relate to studies that focus on the development and validation of outcome measures themselves. When carrying out such studies, it is equally important to choose the appropriate number of participants. However, although an "inappropriate sample size can lead to erroneous findings" (Anthoine et al. 2014: 2), when it comes to development and validation of scales and the identification of appropriate questionnaire structure, there is currently no commonly held standard for sample size as is typical in other clinical research.

Measurement of multifaceted phenomena

As explained above, outcome measures aim to assess the impact of an intervention on specific presenting features or symptoms of patients. Presenting features and symptoms, however, do not exist in isolation. On the contrary, they are embedded in, often complex, contexts; they vary both from person to person and within individuals, they have multiple potential triggers, and they may emerge in diverse ways and within different environments. This reality poses certain challenges when it comes to measuring change in targeted features or symptoms. These challenges have been discussed widely and, in a study regarding back pain, Bolton and Breen comment:

> "Selecting outcome measures for use in research trials in conditions such as back pain […] has always been problematic […] [since pain], the primary symptom of back pain, is a multidimensional, individual experience or behavior with a number of sensory, affective, cognitive/behavioral, and social aspects." (Bolton & Breen 1999: 503)

In some cases, the use of different types of scales in tandem with each other can mitigate issues concerning multidimensionality of presenting features (Fitzpatrick et al. 1998).

It might be the case that similar symptoms arise for different client groups and thus an outcome measure might be transferable in terms of content and presentation. For example, the Immediate and Deferred Prose Memory tests (Novelli et al. 1986) that measure lexical performance and semantic memory have been used with dementia clients in a study exploring a manualised music-based protocol for the rehabilitation of cognitive functions (Ceccato et al. 2012), despite the fact the measure was not specifically developed for this population. Similar tests for memory and lexical performance might also be used for patients with various types of

[6] Statistical power refers to the likelihood that a measurement will distinguish an actual effect from one of chance.

[7] Further considerations regarding sample sizes can be found in the context of randomised controlled trials (Vink, Bruinsma & Scholten 2003), case study research (e.g. Gomm, Hammersley & Foster 2000; Lieberson 1991) and in related music therapy literature (e.g. DeNora & Ansdell, 2014).

trauma, or patients who are undergoing rehabilitation, for instance.

Similar kinds of challenges are faced when exploring how music therapy works in community, medical and other contexts, and when measuring change within such contexts. Here, the difficulties regarding measurement of multifaceted phenomena relates not only to the nature of presenting features outlined earlier, but also to the multifaceted nature of music-making situations which are core to music therapy practice. Reflecting on the difficulties and limitations in developing the IiM scale, for example, Gold et al. (2013) acknowledge that

> "an important conceptual limitation of the IiM scale is that it is organized around various ways of musical engagement (singing, playing, and listening) and not clearly articulating the functional uses of music and the use of music as accompaniment to other activities." (Gold et al. 2013: 678)

The challenges that emerge from the complexities of studying (inter)subjective, multifaceted and contextual phenomena have often been a springboard for debates and critiques of the use of outcome measures in music therapy and of the evidence-based practice movement more generally (Aigen 2015; DeNora & Ansdell 2014). Others suggest an integral understanding of evidence-based music therapy practice (e.g. Abrams 2010; Wheeler & Murphy 2016).

Context specificity and comparability

In addition to the six key considerations discussed above, as well as issues of sample size and measurement of multifaceted phenomena, the selection and use of outcome measures is determined by their context specificity or their comparability. Non-context-specific measures may not be sensitive enough to identify specific details of the phenomenon under study. For this reason, the use of different measures in conjunction with each other has been proposed (Fitzpatrick et al. 1998; Jones, Edwards & Hounsome 2012).

The comparability of measures is connected to cost-effectiveness. Financial resources are distributed partly according to how interventions compare to each other: Which intervention is going to best deliver cost-effective results when implemented? A measure used in isolation does not allow for comparability and, in turn, measuring tools require a reference framework in order to be meaningful. For this reason, choosing a measure that operates within a relevant framework of comparison is arguably just as crucial as choosing one on the basis of its tested validity. Despite the ease of administering and scoring them, the End of Life in Dementia (EOLD) scales, for example, have been critiqued for being valid only for a narrow target group (Parker & Hodgkinson 2011).

Feasibility and psychometric properties

The feasibility and psychometric properties of outcome measures are key issues in the selection, use and usefulness of such measures. Feasibility concerns how straightforward the use and scoring of an outcome measure is. It also relates to considerations around availability, cost and length. Convoluted or long-winded measurement methods can be problematic, particularly when working with vulnerable patient groups where simplicity might be favoured over more thorough measures. The practical feasibility and suitability of a measure for a given context can affect the strength of the collected data (Fitzpatrick et al. 1998).

Psychometric properties of outcome measures relate to the quality and detail of the information generated by the measures. These properties refer to "quantifiable attributes […] that relate to the statistical strength or weakness of a test or measurement" (Medical Dictionary for the Health Professions and Nursing 2012: no pagination). Several outcome measures in music therapy, such as the Music in Dementia Assessment Scales (MiDAS; McDermott et al. 2014) and the MATADOC (Magee et al. 2016), have been assessed for their psychometric properties.

Reliability and validity are two crucial psychometric properties. On the one hand, reliability refers to "the ability of the outcome measure to consistently measure an attribute" (Parker & Hodgkinson 2011: 7). In other words, it refers to the ability of a measure to give consistent results under similar circumstances. Prickett illustrates this with a music-related example:

> "A dependent variable that purported to measure musical aptitude, but which gave widely differing results when administered to the same person three consecutive times or when scored by several different people, would not be reliable, and to attempt to base a study on this measure would be foolish." (Prickett 2005: 54)[8]

[8] A dependent variable is a variable whose value is affected by (i.e. is 'dependent' on) another variable: the independent variable. Assessment typically measures

Reliability assessment tends to depend on numeric tests.[9]

On the other hand, validity is concerned with "the beguilingly simple question of whether a [measure] is truly assessing what it purports to assess" (Fitzpatrick et al. 1998: 2). Some may argue it is an intellectual ideal and an elusive goal that we can never fully reach (Prickett 2005). Assessment of validity tends to relate to the conceptual construction of an outcome measure and relies on close analysis of its items. In other words, validity focuses on the meaning and interpretation of a measure's content.

Tables 1 and 2 outline different types of reliability and validity respectively drawing from several sources: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1985), Cozby (2001), Cronbach (1971), Moskal and Leydens (2000), and Phelan and Wren (2005-2006). A fuller description regarding reliability and validity can be found in Carmines and Zeller (1979), while two examples of studies focusing on the reliability and validity of music therapy-specific measures can be found in Gold et al. (2013) and Magee et al. (2016). These studies concern the IiM and the MATADOC measures respectively.

DISCUSSION: SUMMARY AND REFLECTIONS

In this article we have offered an introduction to terminology around outcome measurement through a music therapy frame. This frame involved not only the consideration of music therapy examples in terms of the application of outcome measurement, but also an outline of current debates regarding outcome measurement in the field. From this point of view, the article does not only introduce outcome measurement terminology, but also hints to the professional discourse around such terminology. Below, after summarising the key terms explored in the article, we reflect on the importance of understanding outcome measurement terminology for future dialogues and developments in the field.

To sum up, an outcome measure is an instrument that is used to assess the effectiveness or impact of an intervention in achieving its aims. This often involves measuring the impact of an intervention on a patient's presenting features or symptoms. Outcome measures may be non-patient-based or patient-based. The former commonly utilise observation, task-based activities or measurement of physiological elements whilst questionnaires and interviews dominate the latter, using the patient as the primary informant. Principal uses of patient-based outcome measures include: healthcare-policy evaluation, healthcare assessment, making clinical decisions in routine practice, economic assessment and resource allocation, clinical audit, as well as monitoring and assessing the health and needs of a population.

In music therapy (and other related disciplines) introductions like the one offered here can bridge gaps between practitioners and researchers as well as professionals from different research traditions who may be less familiar with outcome measurement. This article complements other similar endeavours in music therapy (e.g. Lipe 2015) and beyond.[10] For example, Pasiali, Schoolmeesters and Engen (2016) offer an analysis of resilience-related measures and, after identifying their salient psychometric properties, they draw conclusions about practical uses in music therapy. Also, the online resource *Outcome Measures in Music Therapy* (Cripps, Tsiris & Spiro 2016) gives an overview of existing music therapy-specific outcome measures.

Going back to our opening story, we envisage that the initial understanding of outcome measures offered in this article answers Bob's question: *"Sounds good, but… what is it?"* As a music therapist learns more about outcome measurement and starts using measures in their practice or research, additional questions inevitably emerge, not only in terms of their use and method but also in terms of their fit with different music therapy approaches and theoretical orientations. This is where an understanding of the debates around outcome measurement is informative. As outlined

how changes in the independent variable (e.g. music therapy intervention) cause changes to the dependent one (e.g. musical aptitude).

[9] For more information regarding numerical and statistical approaches, see Meadows (2016), Waldon (2016) and Streiner, Norman and Cairney (2015).

[10] Non-music therapy examples include Kyte et al. (2015) who offer an introduction to patient-reported outcome measures in physiotherapy, as well as Young et al. (2015) who focus on outcome measurement in prosthetics and orthotics.

Type of reliability	Description
Test-retest reliability	This type of reliability assesses whether results are consistently replicable. It can be obtained by administering the same outcome measure twice to the same group of people. The two sets of scores can then be correlated in order to evaluate stability over time.
Parallel forms reliability	This type is obtained by administering to the same group of people different versions of an assessment tool (both versions must contain items that probe the same construct, skill or knowledge base). The scores from the two versions can then be correlated in order to evaluate the consistency of results across alternate versions.
Inter-rater reliability	This type is used to assess the degree to which different raters agree in their measurements. Inter-rater reliability is useful because observers may not interpret material in the same way.
Internal consistency reliability	This type evaluates the degree to which different measure items that probe the same construct produce similar results.
❑ Average inter-item correlation	This subtype of internal consistency reliability is obtained by taking all of the items on a measure that probe the same construct (e.g. reading comprehension), determining the correlation coefficient[11] for each pair of items, and taking the average of all of these correlation coefficients, thus yielding the average inter-item correlation.
❑ Split-half reliability	This subtype of internal consistency reliability starts by splitting in half all items of a measure that are intended to probe the same area of knowledge in order to form two sets of items. The entire measure is administered to a group of individuals, the total score for each set is calculated, and finally the split-half reliability is obtained by determining the correlation between the two total set scores.

Table 1: Types of reliability

Type of validity	Description
Face validity	This type of validity concerns the extent to which the measure is subjectively seen to cover what it purports to assess. Face validity is the type that respondents can easily assess and it may therefore be an essential component in enlisting their motivation. If the respondents do not believe the measure accurately captures their opinions, they may become disengaged with it.
Construct validity	This type is used to ensure that the measure actually tests what it is intended to (i.e. the construct as developed from theory)[12] and not something else. Experts familiar with the construct can examine the items of an outcome measure and decide what each specific item is intended to assess.
Content validity	This type is used to estimate how much a measure represents each element of a construct. This requires expert evaluation of whether the outcome measure items assess what they were intended to assess. For example, in clinical settings, content validity refers to the correspondence between items in the outcome measure and a given set of symptoms.
Criterion validity	This type correlates measurement results with performance or behaviour in another situation. In the other situation a different measure may be used.
Formative validity	This type is used to assess the extent to which a measure can provide information to help improve the intervention under study.
Sampling validity	This type is similar to content validity and ensures that the measure covers the broad range of areas within the construct under investigation. Given that not everything can be covered, items from all of the areas need to be sampled. This may need to be completed by experts to ensure that the content area is adequately sampled.

Table 2: Types of validity

[11] The correlation coefficient gives a statistical relationship between two variables.

[12] In the Music Therapy Coding Scheme (Raglio, Traficante & Oasi, 2006), for example, constructs refer to nonverbal communication, countenance, verbal communication and sonorous musical communication.

earlier, some main concerns pertain to the compartmentalisation of music therapy, the distance from context, the generalisation of results as well as the assumption of cause and effect in music therapy.

Fostering an integral understanding of evidence-based music therapy practice (Abrams 2010; Tsiris et al. 2016) where – instead of antagonism – different research approaches are seen as complementary, we advocate for a critical engagement with outcome measures and their potential uses in music therapy. A respectful understanding of different research terms, methods and orientations necessitates an understanding of their particular contexts of reference. We also argue that reflexivity – although it seems to be discussed more often within qualitative or interpretivist approaches to research (see Wheeler & Murphy 2016; Wheeler & Rickson 2017) – is a necessity for any rigorous enquiry, whether practice- or research-based, and irrespective of its philosophical underpinnings. In our view, reflexivity forms the basis for making balanced claims and fair representations of the results of each enquiry.

From this point of view, and while avoiding epistemological polarities, this article enhances understanding of outcome measures, their characteristics and their uses in music therapy. By offering an introduction to outcome measurement terminology and by giving examples from music therapy, this article also contributes to a more informed engagement with outcome-based research and related debates in the field.

The increased familiarity of music therapists with terminology and procedures involved in outcome-based research is an essential step towards bridging the gap between research and practice, as well as between outcome-based and other types of enquiry; and we argue that music therapy training is well placed to cultivate such familiarity. Likewise, better understanding of outcome measurement leads to a more critical and constructive engagement with such research which seems to be treated, at times, blindly by funders, policy-makers and service providers as the only rigorous approach. However, awareness of the difference, for example, between *efficacy* and *effectiveness* (i.e. between effect under controlled and real-word clinical settings) could help understand how outcome measurement is represented and understood (Fleischhacker & Goodwin 2009; Gartlehner et al. 2006; Wigram & Gold 2012). A study indicating *efficacy* of a music therapy intervention within a particular research context, for example, does not guarantee its *effectiveness* in everyday music therapy contexts.

Outcome measures are ubiquitous in randomised controlled trials, which are, in turn, considered the 'gold standard' in the evidence-based practice movement (Evans 2003; Wigram & Gold 2012). And indeed, such trials in music therapy and other music interventions are growing in number (Kamioka et al. 2014; Mrázová & Celec 2010; Spiro, Tsiris & Pavlicevic 2015; Treurnicht Naylor et al. 2011). Although the philosophical and methodological underpinnings of this type of research (as well as the criteria and assumptions regarding what is considered to be 'robust evidence') have been debated widely both in music therapy (e.g. Abrams 2010; Aigen 2015; Ansdell 2006; DeNora 2006; Stige, Malterud & Midtgarden 2009; Wigram 2006) and in other fields (e.g. Raw et al. 2012; Williams & Garner 2002), such studies play a key role in expanding the current evidence base of music therapy and in shaping new policy initiatives. The exploratory randomised trial by Talwar et al. (2006), for example, contributed to the integration of music therapy in the UK's National Institute for Health and Care Excellence (NICE) guidelines for schizophrenia, while studies including those by Mössler et al. (2011) and Gold et al. (2009) played a role in drawing the attention of policy-makers in Norway and informed the subsequent inclusion of music therapy in the Norwegian Directorate of Health's guidelines for the treatment of psychotic disorders (see Nebelung & Krüger 2015). All these developments, of course, raise a number of possibilities and opportunities as well as dilemmas and risks for music therapy in terms of the identity and quality of music therapy practices, as well as education and professionalisation (Stige 2015).

In closing, we encourage a critical engagement with outcome measurement in music therapy. This requires an understanding of associated terminology, which has been the focus of this article. It also requires an awareness of the debates around outcome measurement and of their implications on the profession and practice of music therapy.

ACKNOWLEDGMENTS

We would like to thank Prof Mercédès Pavlicevic (Nordoff Robbins England and Wales, UK) and Prof Hanne Mette Ridder (Aalborg University, Denmark) for commenting on earlier versions of this article. We also thank Katie Rose Sanfilippo (Nordoff Robbins England and Wales; Goldsmiths,

University of London, UK) for her help in identifying some of the material presented here, and Fiona Crow (Barchester Healthcare; Nordoff Robbins Scotland, UK) for initial feedback.

REFERENCES

Abrams, B. (2010). Evidence-based music therapy practice: An integral understanding. *Journal of Music Therapy, 47*(4), 351-379.

Aigen, K. (2005). *Music-Centered Music Therapy*. Gilsum, NH: Barcelona Publishers.

Aigen, K. (2007). In defense of beauty: A role for the aesthetic in music therapy theory: Part I: The development of aesthetic theory in music therapy. *Nordic Journal of Music Therapy*, *16*(2), 112-128.

Aigen, K. (2008). In defense of beauty: A role for the aesthetic in music therapy theory: Part II: Challenges to aesthetic theory in music therapy: Summary and response. *Nordic Journal of Music Therapy*, *17*(1), 3-18.

Aigen, K. (2015). A critique of evidence-based practice in music therapy. *Music Therapy Perspectives, 33*(1), 12-24.

Albrecht, G. L. (1994). Subjective Health Assessment. In C. Jenkinson (Ed.), *Measuring Health and Medical Outcomes* (pp. 7-26). London: UCL Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Music Therapy Association (2017). *Statement to AMTA members on the TIME-A trial published by Bieleninik et al. (2017)*. Retrieved from: https://www.musictherapy.org/research_and_asd_brief_statement_on_time-a_trial_/

Ansdell, G. (2006). Response to Tia DeNora. *British Journal of Music Therapy, 20*(2), 96-99.

Ansdell, G., & DeNora, T. (2016). *Musical Pathways in Recovery: Community Music Therapy and Mental Wellbeing*. New York, NY: Routledge.

Ansdell, G., & Pavlicevic, M. (2010). Practicing “gentle empiricism” – the Nordoff Robbins research heritage. *Music Therapy Perspectives, 28*(2), 131-139.

Anthoine, E., Moret, L., Regnault, A., Sébille, V., & Hardouin, J. B. (2014). Sample size used to validate a scale: A review of publications on newly-developed patient reported outcomes measures. *Health and Quality of Life Outcomes, 12*. Retrieved from: https://hqlo.biomedcentral.com/articles/10.1186/s12955-014-0176-2

Auer, S. R., Monteiro, I. M., & Reisberg B. (1996). The Empirical Behavioral Pathology in Alzheimer's Disease (E-BEHAVE-AD) Rating Scale. *International Psychogeriatrics, 8*(2), 247-266.

Barr, J. (1995). The outcomes movement and health status measures. *Journal of Allied Health, 24*(1), 13-28.

Bausewein, C., Daveson, B., Benalia, H., Simon, S., & Higginson, I. (2011). *Outcome Measurement in Palliative Care: The Essentials.* London: PRISMA.

Bieleninik, Ł., Geretsegger, M., Mössler, K., Assmus, J., Thompson, G., Gattino, G., Elefant, C., Gottfried, T., Igliozzi, R., Muratori, F., Suvini, F., Kim, J., Crawford, M., Odell-Miller, H., Oldfield, A., Casey, Ó., Finnemann, J., Carpente, J., Park, A-l., Grossi, E., & Gold, C. (2017). Effects of improvisational music therapy vs enhanced standard care on symptom severity among children with autism spectrum disorder: The TIME-A randomized clinical trial. *Journal of the American Medical Association, 318*(6), 525-535. Retrieved from: https://jamanetwork.com/journals/jama/article-abstract/2647867

Bolton, J. E., & Breen, A. C. (1999). The Bournemouth questionnaire: A short-form comprehensive outcome measure. I. Psychometric properties in back pain patients. *Journal of Manipulative and Physiological Therapeutics*, *22*(8), 503-509.

Bruscia, K. (1987). *Improvisational Models of Music Therapy.* Springfield, IL: Charles C Thomas.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and Validity Assessment*. London: Sage.

Ceccato, E., Vigato, G., Bonetto, C., Bevilacqua, A., Pizziolo, P., Crociani, S., Zanfretta, E., Pollini, L., Caneva, A., Baldin, L., Frongillo, C., Signorini, A., Demoro, S., & Barchi, E. (2012). STAM protocol in dementia: A multicenter, single-blind, randomized, and controlled trial. *American Journal of Alzheimer's Disease and Other Dementias, 27*(5), 301-310.

The Chartered Society of Physiotherapy (2013). *Patient Reported Outcome Measures*. Retrieved from: http://www.csp.org.uk/tagged/patient-reported-outcome-measures-proms-0

Chu, H., Yang, C. Y., Lin, Y., Ou, K. L., Lee, T. Y., O’Brien, A. P., & Chou, K. R. (2013). The impact of group music therapy on depression and cognition in elderly persons with dementia: A randomized controlled study. *Biological Research for Nursing*, *16*(2), 209-217.

Collin, C., Wade, D. T., Davies, S., & Horne, V. (1988). The Barthel ADL Index: A reliability study. *International Disability Studies. 10*(2), 61-63.

Cozby, P.C. (2001). *Methods in Behavioral Research* (7th Edition). California, CA: Mayfield Publishing Company.

Cripps, C., Tsiris, G., & Spiro, N. (Eds.). (2016). *Outcome measures in music therapy: A resource developed by the Nordoff Robbins research team*. London: Nordoff Robbins. Retrieved from: https://www.nordoff-robbins.org.uk/ResearchResources

Cronbach, L. J. (1971). Test Validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd Edition). Washington, DC: American Council on Education.

De Backer, J., & Sutton, J. (Eds.). (2014). *The Music in Music Therapy: Psychodynamic Music Therapy in Europe: Clinical, Theoretical and Research Approaches*. London: Jessica Kingsley Publishers.

DeNora, T. (2006). Evidence and effectiveness in music therapy: Problems, possibilities and performance in health contexts. *British Journal of Music Therapy, 20*(2), 81-99.

DeNora, T., & Ansdell, G. (2014). What can't music do? *Psychology of Well-Being, 4*, 2-10. Retrieved from: https://link.springer.com/article/10.1186/s13612-014-0023-6

Donabedian, A. (1966). Evaluating the quality of medical care. *The Milbank Memorial Fund Quarterly*, *44*(3), 166-206.

Ebrahim, S. (1995). Clinical and public health perspectives and applications of health related quality of life measurement. *Social Science & Medicine, 41*, 1383-1394.

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*, *12*(1), 77-84.

Farrant, C., Tsiris, G., & Pavlicevic, M. (2014). *A Guide to Research Ethics for Arts Therapists and Arts & Health Practitioners*. London: Jessica Kingsley Publishers.

Fitzpatrick, R., Davey, C., Buxton, M. J., & Jones, D. R. (1998). Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment, 2*(14), 1-75.

Fleischhacker, W.W., & Goodwin, G.M. (2009). Effectiveness as an outcome measure for treatment trials in psychiatry. *World Psychiatry, 8*(1), 23-27. Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/j.2051-5545.2009.tb00200.x/full

Garland, A. F., Kruse, M., & Aarons, G. A. (2003). Clinicians and outcome measurement: What's the use? *The Journal of Behavioral Health Services and Research, 30*(4), 393-405.

Gartlehner, G., Hansen, R. A., Nissman, D., Lohr, K. N., & Carey, T. S. (2006). *Criteria for distinguishing effectiveness from efficacy trials in systematic reviews*. Rockville, MD: Agency for Healthcare Research and Quality.

Gilbody, S. M., House, A. O., & Sheldon, T. A. (2003). *Outcome Measures in Psychiatry: A Critical Review of Outcomes Measurement in Psychiatric Research and Practice*. York: York Publishing Services Ltd.

Gold, C., & Bieleninik, Ł. (2018). Authors' response. *Nordic Journal of Music Therapy, 27*(1), 90-92.

Gold, C., Rolvsjord, R., Mössler, K., & Stige, B. (2013). Reliability and validity of a scale to measure interest in music among clients in mental health care. *Psychology of Music*, *41*(5), 665-682.

Gold, C., Solli, H. P., Krüger, V., & Lie, S. A. (2009). Dose–response relationship in music therapy for people with serious mental disorders: Systematic review and meta-analysis. *Clinical Psychology Review*, *29*(3), 193-207.

Gomm, R., Hammersley, M., & Foster, P. (Eds.). (2000). *Case Study Method: Key Issues, Key Texts*. London: Sage.

Grant, R.E. (1995). Music Therapy Assessment for Developmentally Disabled Clients. In T. Wigram, B. Saperston & R. West (Eds.), *The Art and Science of Music Therapy: A Handbook* (pp.273-287). London: Routledge.

Guo, J. H., Chen, H. J., & Luh, W. M. (2011). Sample size planning with the cost constraint for testing superiority and equivalence of two independent groups. *British Journal of Mathematical and Statistical Psychology, 64*, 439-461.

Hackett, S. (2016). The combined arts therapies team: Sharing practice development in the National Health Service in England. *Approaches: An Interdisciplinary Journal of Music Therapy, Special Issue 8*(1), 42-49. Retrieved from: http://approaches.gr/el/hackett-a20160109/

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8). Retrieved from: http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

Jones, C., Edwards, R. T., & Hounsome, B. (2012). Health economics research into supporting carers of people with dementia: A systematic review of outcome measures. *Health and Quality of Life Outcomes, 10*(142), 1-8.

Kamioka, H., Tsutani, K., Yamada, M., Park, H., Okuizumi, H., Tsuruoka, K., Honda, T., Okada, S., Park, S. J., Kitayuguchi, J., Abe, T., Handa, S., Oshio, T., & Mutoh, Y. (2014). Effectiveness of music therapy: A summary of systematic reviews based on

randomized controlled trials of music interventions. *Patient Preference and Adherence*, *8*, 727-754.

Kyte, D. G., Calvert, M., van der Wees, P. J., Ten Hove, R., Tolan, S., & Hill, J. C. (2015). An introduction to patient-reported outcome measures (PROMs) in physiotherapy. *Physiotherapy, 101*(2), 119-125.

Langan, D. (2009). A music therapy assessment tool for special education: Incorporating education outcomes. *Australian Journal of Music Therapy, 20*, 78-98.

Lieberson, S. (1991). Small N's and big conclusions: An examination of the reasoning in comparative studies based on a small number of cases. *Social Forces, 70*(2), 307-320.

Lipe, A. (2015). Music Therapy Assessment. In B. Wheeler (Ed.), *Music Therapy Handbook* (pp. 76-90). New York, NY: The Guildford Press.

Lyons, J. S., Howard, K., O'Mahoney, M., & Lish, J. (1997). *The Measurement and Management of Clinical Outcomes in Mental Health*. New York, NY: John Wiley & Sons.

Magee, W. L. (2007). Development of a music therapy assessment tool for patients in low awareness states. *NeuroRehabilitation, 22*(4), 319-324.

Magee, W. L., Siegert, R. J., Taylor, S. M., Daveson, B. A., & Lenton-Smith, G. (2016). Music Therapy Assessment Tool for Awareness in Disorders of Consciousness (MATADOC): Reliability and validity of a measure to assess awareness in patients with disorders of consciousness. *Journal of Music Therapy, 53*(1), 1-26.

Mahoney, J. F. (2010). Interrater agreement on the Nordoff-Robbins evaluation scale I: Client-therapist relationship in musical activity. *Music and Medicine*, *2*(1), 23-28.

Maue-Johnson, E.L., & Tanguay, C.L. (2006). Assessing the unique needs of hospice patients: A tool for music therapists. *Music Therapy Perspectives, 24*(1), 13-20.

McDaniel, R. W., & Bach, C. A. (1994). Quality of life: A concept. *Nursing Research, 3*, 18-22.

McDermott, O., Orgeta, V., Ridder, H. M., & Orrell, M. (2014). A preliminary psychometric evaluation of Music in Dementia Assessment Scales (MiDAS). *International Psychogeriatrics, 26*(06), 1011-1019.

McDowell, I., & Newell, C. (1996). *Measuring Health: A Guide to Rating Scales and Questionnaires*. Oxford: Oxford University Press.

Meadows, A. (2016). Introduction to Statistical Concepts. In B.L. Wheeler & K. Murphy (Eds.), *Music Therapy Research* (3rd Edition, Chapter 18). Dallas, TX: Barcelona Publishers.

Medical Dictionary for the Health Professions and Nursing (2012). *Psychometric properties*. Retrieved from http://medical-dictionary.thefreedictionary.com

Michener, L. A. (2011). Patient-and clinician-rated outcome measures for clinical decision making in rehabilitation. *Journal of Sport Rehabilitation, 20*(1), 37-45.

Moskal, B.M., & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10). Retrieved from: http://pareonline.net/getvn.asp?v=7&n=10

Mössler, K., Chen, X., Heldal, T. O., & Gold, C. (2011). *Music therapy for people with schizophrenia and schizophrenia-like disorders*. The Cochrane Library. Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004025.pub3/full

Mrázová, M., & Celec, P. (2010). A systematic review of randomized controlled trials using music therapy for children. *Journal of Alternative and Complementary Medicine, 16*(10), 1089-1095.

Nebelung, I., & Krüger, V. (2015). Norway: Country report on professional recognition of music therapy. *Approaches: Music Therapy & Special Music Education, Special Issue 7*(1), 171-172. Retrieved from: http://approaches.gr/special-issue-7-1-2015/

Nordoff, P., & Robbins, C. (1977). *Creative Music Therapy*. New York, NY: John Day.

Nordoff, P., & Robbins, C. (2007). *Creative Music Therapy: A Guide to Fostering Clinical Musicianship* (Revised Edition). Gilsum, NH: Barcelona Publishers.

Novelli, G., Papagno, C., Capitani, E., Laiacona, M., Vallar, G., & Cappa, S. F. (1986). Tre test clinici di ricerca e produzione lessicale. Taratura su soggetti normali. *Archivio di Psicologia, Neurologia e Psichiatria, 47*(4), 477-506.

Oldfield, A. (2006). *Interactive Music Therapy in Child and Family Psychiatry: Clinical Practice, Research, and Teaching*. London: Jessica Kingsley Publishers.

Parker, D., & Hodgkinson, B. (2011). A comparison of palliative care outcome measures used to assess the quality of palliative care provided in long-term care facilities: A systematic review. *Palliative Medicine, 25*(1), 5-20.

Pasiali, V., Schoolmeesters, L., & Engen, R. (2016). Mapping resilience: Analyses of measures and suggested uses in music therapy. *Approaches: An Interdisciplinary Journal of Music Therapy*, First View, 1-25. Retrieved from: http://approaches.gr/pasiali-a20160920/

Pavlicevic, M., & Ansdell, G. (Eds.). (2004). *Community Music Therapy*. London: Jessica Kingsley.

Phelan C., & Wren, J. (2005-2006). *Exploring reliability in academic assessment*. Retrieved from: https://www.uni.edu/chfasoa/reliabilityandvalidity.htm

Prickett, C. (2005). Principles of Quantitative Research. In B. Wheeler (Ed.), *Music Therapy Research*

(pp. 45-58). Gilsum, NH: Barcelona Publishers.

Procter, S. (2011). Reparative musicing: Thinking on the usefulness of social capital theory within music therapy. *Nordic Journal of Music Therapy, 20*(3), 242-262.

Procter, S. (2018). Addicted to the RCT? Reasons for the Epidemic and Routes to Recovery. In C. Warner, G. Tsiris & T. Watson (Eds.), *Music, Diversity and Wholeness (Book of Abstracts, Third BAMT Conference, 16-18 February 2018)* (p. 218). London: British Association for Music Therapy.

Raglio, A., Bellelli, G., Traficante, D., Gianotti, M., Ubezio, M. C., Gentile, S., Villani, D., & Trabucchi, M. (2010). Efficacy of music therapy treatment based on cycles of sessions: A randomised controlled trial. *Aging and Mental Health, 14*(8), 900-904.

Raglio, A., Traficante, D., & Oasi, O. (2006). A coding scheme for the evaluation of the relationship in music therapy sessions. *Psychological Reports, 99*(1), 85-90.

Raglio, A., Traficante, D., & Oasi, O. (2007). Comparison of the music therapy coding scheme with the music therapy checklist. *Psychological Reports, 101,* 875-880.

Raw, A., Lewis, S., Russell, A., & Macnaughton, J. (2012). A hole in the heart: Confronting the drive for evidence-based impact research in arts and health. *Arts & Health, 4*(2), 97-108.

Ridder, H M., & Tsiris, G. (Eds.). (2015). Special issue on 'Music therapy in Europe: Paths of professional development". *Approaches: Music Therapy & Special Music Education, 7*(1). Retrieved from: http://approaches.gr/special-issue-7-1-2015/

Ryan, C., & Charragain, C. N. (2010). Teaching emotion recognition skills to children with autism. *Journal of Autism & Developmental Disorders, 40*(12), 1505-1511.

Schwartzberg, E. T., & Silverman, M. J. (2013). Effects of music-based social stories on comprehension and generalization of social skills in children with autism spectrum disorders: A randomized effectiveness study. *The Arts in Psychotherapy, 40*(3), 331-337.

Solomon, P. R., & Pendlebury, W. W. (1998). Recognition of Alzheimer's disease: The 7 Minute Screen. *Family Medicine, 30*(4), 265-271.

Spiro, N., Farrant, C., Himberg, T., & Pavlicevic, M. (2016). What Do Music Therapists Look for When Assessing Clients in Music Therapy Sessions? In U. Aravinth, M. Pavlicevic & G. Watts (Eds.), *Re-Visioning our Voice: Resourcing Music Therapy for Contemporary Needs (Book of Abstracts, Second BAMT Conference, 8-10 April 2016)* (p. 139). London: British Association for Music Therapy.

Spiro, N., Tsiris, G., & Cripps, C. (2018). A systematic review of outcome measures in music therapy. *Music Therapy Perspectives,* 36(1), 67-78.

Spiro, N., Tsiris, G., & Pavlicevic, M. (2014). Music Therapy Models. In W. F. Thompson (Ed.), *Music in the Social and Behavioral Sciences: An Encyclopedia* (pp. 771–773). Thousand Oaks: Sage.

Spiro, N., Tsiris, G., & Pavlicevic, M. (2015). Music Practices in Dementia Care: A critical Overview of Randomised Controlled Trials. In *Music Therapy across Contexts (Book of Abstracts, 8th Nordic Music Therapy Congress, 4-8 August 2015)* (p. 91). Oslo: Norsk Forening for Musikkterapi.

Stegemann, T., Schmidt, H. U., Fitzthum, E., & Timmermann, T. (Eds.). (2016). *Music Therapy Training Programmes in Europe: Theme and Variations*. Wiesbaden: Reichert Verlag.

Stige, B. (2002). *Culture-Centered Music Therapy*. Gilsum, NH: Barcelona Publishers.

Stige, B. (2015). *The context-renewing collaborative doing of music therapy: Towards a philosophy of professional practice*. Keynote presentation at the 8th Nordic Music Therapy Congress "Music Therapy Across Contexts", 4-8 August 2015, Oslo, Norway.

Stige, B., Ansdell, G., Elefant, C., & Pavlicevic, M. (2010). *Where Music Helps: Community Music Therapy in Action and Reflection*. Aldershot: Ashgate.

Stige, B., Malterud, K., & Midtgarden, T. (2009). Toward an agenda for evaluation of qualitative research. *Qualitative Health Research, 19*(10), 1504-1516.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health Measurement Scales: A Practical Guide to their Development and Use* (5th Edition). Oxford: Oxford University Press.

Talwar, N., Crawford, M., Maratos, A., Nur, U., McDermott, O., & Procter, S. (2006). Music therapy for in-patients with schizophrenia: Exploratory randomized trial. *The British Journal of Psychiatry, 189*, 405-409.

Testa, M. A., & Nackley, J. F. (1994). Methods for quality-of-life studies. *Annual Review of Public Health, 15*, 535-659.

Trauer, T. (Ed.). (2010). *Outcome Measurement in Mental Health: Theory and Practice*. Cambridge: Cambridge University Press.

Treurnicht Naylor, K., Kingsnorth, S., Lamont, A., McKeever, P., & Macarthur, C. (2011). The effectiveness of music in pediatric healthcare: A systematic review of randomized controlled trials. *Evidence-Based Complementary and Alternative Medicine,* 1-18. Retrieved from: https://www.hindawi.com/journals/ecam/2011/464759/abs/

Trondalen, G., & Bonde, L. O. (2012). Music Therapy: Models and Interventions. In R. MacDonald, G. Kreutz & L. Mitchell (2012). *Music, Health, and*

Wellbeing (pp. 40-62). Oxford: Oxford University Press.

Tsiris, G. (2008). Aesthetic experience and transformation in music therapy: A critical essay. *Voices: A World Forum for Music Therapy, 8*(3). Retrieved from: https://www.voices.no/index.php/voices/article/view/416

Tsiris, G., Derrington, P., Sparkes, P., Spiro, N., & Wilson, G. (2016). Interdisciplinary dialogues in music, health and wellbeing: Difficulties, challenges and pitfalls. In M. Belgrave (Ed.), *Proceedings of the ISME Commission on Special Music Education and Music Therapy* (20-23 July 2016, Edinburgh, Scotland) (pp. 58-70). Edinburgh: ISME. Retrieved from: https://www.isme.org/other-publications/isme-commission-special-education-and-music-therapy-2016

Tsiris, G., Spiro, N., & Pavlicevic, M. (2018). Repositioning music therapy service evaluation: A case of five Nordoff-Robbins music therapy service evaluations in neuro-rehabilitation. *Nordic Journal of Music Therapy, 27*(1), 3-27.

Turry, A. (2018a). Response to effects of improvisational music therapy vs. enhanced standard care on symptom severity among children with autism spectrum disorder: The TIME-A randomized clinical trial. *Nordic Journal of Music Therapy, 27*(1), 87-89.

Turry, A. (2018b). Actually, music therapy does work. *Music & Medicine, 10*(2), 113-114.

Vink, A.C., Bruinsma, M.S., & Scholten R.J.P.M. (2003). *Music therapy for people with dementia*. The Cochrane Library. Issue 9. Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD003477.pub2/full

Waldon, E. G. (2016). Overview of Measurement Issues in Objectivist Research. In B.L. Wheeler & K. Murphy (Eds.), *Music Therapy Research* (3rd Edition, Chapter 14). Dallas, TX: Barcelona Publishers.

Ware, J. E. (1995). The status of health assessment in 1994. *Annual Review of Public Health, 16*, 327-354.

Wells, N.F. (1988). An individual music therapy assessment procedure for emotionally disturbed young adolescents. *The Arts in Psychotherapy, 15*, 47-54.

Wheeler, B.L., & Murphy, K. (Eds.). (2016). *Music Therapy Research* (3rd Edition). Dallas, TX: Barcelona Publishers.

Wheeler, B., & Rickson, D. (2017). The third edition of 'Music Therapy Research': An interview with Barbara Wheeler. *Approaches: An Interdisciplinary Journal of Music Therapy,* First View, 1-5. Retrieved from: http://approaches.gr/wheeler-i20170118/

Wigram, T. (1999). Assessment methods in music therapy: A humanistic or natural science framework. *Nordic Journal of Music Therapy, 8*(1), 6-24.

Wigram, T. (2006). Response to Tia DeNora. *British Journal of Music Therapy, 20*(2), 93-96.

Wigram, T., & Gold, C. (2012). The Religion of Evidence-Based Practice: Helpful or Harmful to Health and Wellbeing? In R. MacDonald, G. Kreutz & L. Mitchell (Eds.), *Music, Health, and Wellbeing* (pp. 164-182). Oxford: Oxford University Press.

Williams, D. D. R., & Garner, J. (2002). The case against 'the evidence': A different perspective on evidence-based medicine. *The British Journal of Psychiatry*, *180*(1), 8-12.

Wilson, P. (2017). Why didn't music therapy help autistic kids? Maybe the researchers failed, not the therapy. *MedPage Today.* Retrieved from: https://www.medpagetoday.com/pediatrics/autism/67136

Wood, S. (2015). *The performance of community music therapy evaluation*. PhD Thesis, Nordoff Robbins / City University London, UK.

Wood, S. (2016). *A Matrix for Community Music Therapy Practice*. Gilsum, NH: Barcelona Publishers.

Young, J., Rowley, L., Lalor, S., Cody, C., & Woolley, H. (2015). *Measuring Change: An Introduction to Clinical Outcome Measures in Prosthetics and Orthotics*. London: British Association of Prosthetists and Orthotists.

Suggested citation:

Spiro, N., Tsiris, G., & Cripps, C. (2018). *"Sounds good, but… what is it?"* An introduction to outcome measurement from a music therapy perspective. *Approaches: An Interdisciplinary Journal of Music Therapy,* First View (Advance online publication), 1-18.