

**UCSF**

**UC San Francisco Previously Published Works**

**Title**

The Reliability of Rating via Audio-Recording Using the Mindfulness-Based Interventions: Teaching Assessment Criteria.

**Permalink**

<https://escholarship.org/uc/item/3zm053j7>

**Authors**

Floyd, Erin  
Adler, Shelley R  
Crane, Rebecca S  
[et al.](#)

**Publication Date**

2023

**DOI**

10.1177/27536130221149966

Peer reviewed

# The Reliability of Rating via Audio-Recording Using the Mindfulness-Based Interventions: Teaching Assessment Criteria

Global Advances in Integrative Medicine and Health

Volume 12: 1–10

© The Author(s) 2023





Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/27536130221149966

[journals.sagepub.com/home/gam](https://journals.sagepub.com/home/gam)



Erin Floyd, MD<sup>1,2</sup> , Shelley R. Adler, PhD<sup>1</sup>, Rebecca S. Crane, PhD<sup>3</sup> , Judson Brewer, MD, PhD<sup>4</sup>, Patricia Moran, PhD<sup>1</sup>, Robert Richler, MA<sup>1</sup>, Wendy Hartogensis, PhD<sup>1</sup>, Willem Kuyken, PhD<sup>5</sup> , and Frederick M. Hecht, MD<sup>1</sup> 

## Abstract

**Background:** The Mindfulness-Based Interventions: Teaching Assessment Criteria (MBI:TAC) is an important tool for assessing teacher skill and aspects of the fidelity of mindfulness-based interventions, but prior research on and implementation of the MBI:TAC has used video recordings, which can be difficult to obtain, share for assessments, and which increase privacy concerns for participants. Audio-only recordings might be a useful alternative, but their reliability is unknown.

**Objective:** To assess evaluator perception of the rating process and inter-rater reliability of MBI:TAC ratings using audio-only recordings.

**Methods:** We prepared audio-only files from video recordings of 21 previously rated Mindfulness-Based Stress Reduction teachers. Each audio recording was rated by 3 trained MBI:TAC assessors drawn from a pool of 12 who had previously participated in rating the video recordings. Teachers were rated by evaluators who had not viewed the video recording and did not know the teacher. We then conducted semi-structured interviews with evaluators.

**Results:** On the 6 MBI:TAC domains, the intraclass correlation coefficients (ICCs) for audio recordings ranged from .53 to .69 using an average across 3 evaluators. Using a single rating resulted in lower ICCs (.27-.38). Bland-Altman plots showed audio ratings had little consistent bias compared to video recordings and agreed more closely for teachers with higher ratings. Qualitative analysis identified 3 themes: video recordings were particularly helpful when rating less skillful teachers, video recordings tended to provide a more complete picture for rating, and audio rating had some positive features.

**Conclusions:** Inter-rater reliability of the MBI:TAC using audio-only recordings was adequate for many research and clinical purposes, and reliability is improved when using an average across several evaluators. Ratings using audio-only recordings may be more challenging when rating less experienced teachers.

## Keywords

MBSR, MBI:TAC, mindfulness, intervention fidelity, intervention integrity

Received April 23, 2023; Revised September 21, 2023. Accepted for publication November 7, 2023

<sup>1</sup>Osher Center for Integrative Medicine, University of California, San Francisco, CA, USA

<sup>2</sup>Department of Medicine, University of Wisconsin Hospital and Clinics, Madison, WI, USA

<sup>3</sup>Centre for Mindfulness Research and Practice, Bangor University, Bangor, UK

<sup>4</sup>Mindfulness Center at Brown, Brown University, Providence, RI, USA

<sup>5</sup>Oxford Mindfulness Centre, Oxford University, Oxford, UK

## Corresponding Author:

Frederick M. Hecht, MD, Osher Center for Integrative Health, UCSF, UCSF Box 1726, San Francisco, CA 94143-1726, USA.

Email: [rick.hecht@ucsf.edu](mailto:rick.hecht@ucsf.edu)



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and

Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

## Introduction

Mindfulness-based interventions (MBIs) are increasingly being offered in healthcare, education, and community settings. The seminal program, Mindfulness-Based Stress Reduction (MBSR), developed by Jon Kabat-Zinn at University of Massachusetts Medical School, has trained almost 1000 MBSR teachers across the United States and in more than 30 countries. There is a growing scientific literature supporting efficacy for conditions such as pain,<sup>1,2</sup> stress, and anxiety.<sup>3</sup> Related interventions, such as Mindfulness-based Cognitive Therapy (MBCT) have been shown to be effective for depression.<sup>4,5</sup>

In contrast to pharmacologic treatments that can readily be manufactured to ensure consistent active ingredients, MBIs are complex, multi-dimensional interventions, making them challenging to implement to agreed standards for practice.<sup>6,7</sup> A dearth of ways to evaluate the fidelity of such complex interventions, especially teacher skill, has been an important limitation in the field.<sup>6</sup>

The MBI:Teaching Assessment Criteria (MBI:TAC) was developed to measure teaching competency and is now used in both MBI research and training settings.<sup>8-11</sup> An important logistical challenge with the use of the MBI:TAC in research is that it was developed and validated using video (with audio) recordings, because the developers of the tool thought that videos were more informative and, therefore, preferable to audio-only samples.<sup>8</sup> This idea was based on the premise that a core teaching methodology in MBIs is communication of mindfulness through the teacher's embodied practice. Much of this is sensed by the course participants through the body language of the teacher. Video recordings, however, create important logistical challenges, which include more complex requirements for recording (video camera or smartphone with tripod vs smartphone or small audio recorder), greater intrusiveness in the teaching setting due to the more visible equipment, and greater loss of privacy for participants if their faces appear in the video. In addition, video files for two-hour classes are large, adding complexity to storage and transfer of files if needed for rating purposes. The greater visibility of video recording equipment may also increase Hawthorne effects (the alteration of behavior by the subjects of a study due to their awareness of being observed). Video recordings may also increase the possibility of introducing implicit biases based on visual impressions that could influence ratings. As MBI delivery is increasingly conducted online, some of the drawbacks of video-recordings have been reduced; for example, video-conference platforms can make video-recording sessions easy. Even for programs that are delivered using a video-conference platform, however, some of the limitations of using video recordings remain, including privacy concerns for participants and the resulting large files, which are more difficult to store and transfer securely when shared with evaluators.

Audio recordings may be an important alternative to video recordings for assessing teacher skill in some settings, but the MBI:TAC has not been validated using audio-only

recordings. We sought to evaluate the reliability of the MBI:TAC when audio-only recordings were used. Using a mixed-methods approach, we investigated whether the recording format of the MBI sessions influenced the inter-rater reliability of the MBI:TAC and explored MBI:TAC evaluators' perceptions of rating using audio-only recordings. We hypothesized that inter-rater reliability, as measured by intraclass correlation (ICC) coefficients, would be lower with audio recordings than video recordings, though potentially still adequate for research settings.

## Methods

We developed an audio-ratings sub-study within the Predictors of Outcomes in MBSR Participants from Teacher Factors (PrOMPT) trial. This study was reviewed and approved by the Institutional Review Board of University of California, San Francisco. For the PrOMPT-F study, we conducted an 8-week course of 2-hour weekly sessions to train 31 experienced MBI teachers in using the MBI:TAC. The MBI evaluators who conducted MBI:TAC ratings for research purposes had at least 3 years of MBI teaching experience. Trainees were asked to complete weekly homework ratings during the training as well as rate a set of selected video clips at the end of the training. From this pool of newly trained evaluators, we assembled a group of 19 who had both high reliability of ratings compared to benchmark ratings and time available to complete further ratings of video recordings of MBSR teachers. These same evaluators were subsequently invited to participate in the sub-study of MBI:TAC rating using audio recordings. Twelve evaluators agreed to perform ratings for the audio rating study and completed MBI:TAC ratings for at least 1 audio recording.

### *MBSR Course Recordings*

For the main PrOMPT study, 21 teachers recruited from 5 different sites video-recorded themselves teaching MBSR. When using the term "video," we are designating a video recording that includes an audio track. We used a random number generator to select 2 recordings from each teacher for rating, 1 session from the first 4 weeks of the course and a second random selection of a session from the second 4 weeks. There were 40 MBSR session recordings (2 of each teacher, except for 2 teachers who each had just 1 recorded session). For this sub-study, we used only the audio portion of the video recordings that had previously been rated in the main study.

### *MBI:TAC Measure and Ratings*

The MBI:TAC is used to assess the competence and adherence of MBI teaching practice. Evaluators score each domain on a scale from 1 (incompetent) to 6 (advanced).<sup>10</sup> The 6 different domains are: (1) coverage, pacing and

organization of session curriculum, (2) relational skills, (3) embodiment of mindfulness, (4) guiding mindfulness practices, (5) conveying course themes through interactive inquiry and didactic teaching, and (6) holding the group learning environment.

Each MBSR audio-recorded session was rated by 3 different evaluators. Evaluators were assigned to audio recordings for teachers who were unknown to them and for whom they had not already rated using video recordings.

### Quantitative Analysis

For the primary analysis, we calculated the absolute agreement intraclass correlation (ICC) coefficients to assess inter-rater reliability for audio ratings. In this context, ICC is a measure of the agreement between ratings made by multiple evaluators measuring the same MBSR teacher, where 0 indicates no agreement between evaluators, and 1.0 indicates perfect agreement, and the evaluators are considered a random sample from a pool of possible evaluators. We calculated ICC for the audio recording ratings 2 different ways for the 6 MBI:TAC domains, based on absolute agreement, from 2-way random effect models. We calculated individual rater ICC coefficients, which generalize to the case of using a single rater to evaluate a teacher. From the same mixed-effects model, we also calculated ICCs for the average rating of the 3 evaluators. This generalizes to the case of using a panel of evaluators (eg, a panel of 3 evaluators) and averaging their ratings to derive a final rating. In additional analyses, we calculated ICCs comparing inter-rater reliability of audio ratings to those of video ratings. We used paired t-tests to assess whether there were statistically significant differences between the ratings of audio or video recordings of the same teacher. We also used a Bland-Altman plot to evaluate degree of agreement between ratings of audio and video recordings and whether there were generally higher or lower ratings using the audio recordings compared to video.<sup>12</sup> Lastly, we also evaluated whether experienced MBSR teachers were easier to rate using audio alone compared to less experienced teachers using a linear mixed model with teacher years of formal practice or MBSR teaching as predictor and a rating by evaluators of how hard it was to assess an MBSR teacher with only audio. The rating scale ranged from 1 to 5, (higher numbers = harder to rate audio, lower numbers = easier to rate audio), with crossed random effects of teacher and rater (because this was a crossed design in which every category of 1 factor co-occurred in the design with every category of the other factor). For this analysis, data from eleven evaluators were used since 1 rater was not able to complete the survey assessing difficulty or ease of rating MBI:TAC when using video vs audio-recorded sessions.

### Qualitative Analysis

We individually interviewed 8 MBI:TAC evaluators to assess their experience rating sessions using both recording formats. The evaluators we interviewed were a convenience sample

based on who was available and willing to be interviewed; they represented two-thirds of the evaluators. Evaluators received a \$30 gift card in appreciation of the time spent being interviewed. We used a semi-structured interview guide that included questions on evaluators' overall opinions of the MBI:TAC, what they found to be the easiest and most difficult aspects of rating MBSR sessions using both the audio and video recording formats, and how the experience of rating influenced assessors' training and teaching. We conducted the 30-minute interviews in English through a recorded videoconference. Participants were not paid for participating in the interview but received monetary compensation for each MBI:TAC audio rating assignment they completed. Interviews were transcribed verbatim and uploaded to Dedoose (v8.2.14, 2019) for analysis. We conducted qualitative thematic analysis of interviews using an inductive approach. Two team members (RR and EF) independently coded transcripts and jointly reconciled coding differences.<sup>13</sup> The full team met regularly during the coding and analysis process to review coding and to identify and reach consensus on the development of key themes.

## Results

We analyzed MBI:TAC ratings of audio recordings of 40 MBSR sessions from 21 teachers that had previously been rated using video recordings. These 21 teachers were rated by 12 evaluators. The MBI:TAC evaluators had an average of just over 10 years of experience teaching MBIs (Table 1). The MBSR teachers being rated had a range of experience teaching MBIs, from 1 to 33 years. Each teacher was rated by 3 evaluators, except for 1 teacher who had a single audio rating by a fourth rater. The range of teachers rated by each evaluator was 1-11, with a mean of 5.3 teachers rated by each evaluator.

### Quantitative Analysis

For the 6 MBI:TAC domains, individual rater ICC coefficients (which generalize to the case of using a single rater), ranged from .27 to .43 (Table 2). When ICCs were calculated using the average of 3 evaluators (which generalize to the case of using a panel of 3 evaluators), ICCs improved substantially, to a range of .53-.69. ICCs for audio ratings were highest for the domain of guiding mindfulness practice, lowest for the domain of holding the group environment.

When we compared the average final ratings of audio recordings to video recordings, we found that ratings on each of the 6 domains of the MBI:TAC were lower when audio recordings were rated, with *P*-values of <.005 for every domain and average differences that ranged from .13 to .43 on the different domains (Table 3). We used Bland-Altman plots to further compare the agreement between audio and video ratings using the MBI:TAC (Figure 1). These confirmed a bias toward lower ratings using audio recordings, although

**Table 1.** Characteristics of MBI:TAC Evaluators and MBSR Teachers.

Variable	MBI:TAC Evaluators (n = 12)	MBSR Teachers (n = 21)
Age (in years), mean (SD)	55.3, (7.0)	58.7 (10.2)
Race, % (n)	100% caucasian (12)	95.2% caucasian (20) 4.8% biracial (1) – (Asian/European)
Ethnicity, % (n)	8.3% hispanic (1) 91.7% non-hispanic (11)	4.8% hispanic (1) 95.2% non-hispanic (20)
Female, % (n)	91.7% (11)	81.0% (17)
Years of mindfulness personal practice, mean (SD) [range]	19.2 (7.2) [10, 30]	20.1 (11.4) [3, 40]
Years of mindfulness teaching experience, mean (SD) [range]	10.9 (4.3) [7, 22]	8.9 (8.2) [1, 33]

**Table 2.** Intraclass Correlation Coefficients (ICC) for MBI:TAC Audio Ratings by Domain.

Domain	Domain Name	Measurement Type	ICC
1	Coverage pacing and organization of session curriculum	Individual	.38
		Average	.65
2	Relational skills	Individual	.28
		Average	.54
3	Embodiment of mindfulness	Individual	.29
		Average	.55
4	Guiding mindfulness practices	Individual	.43
		Average	.69
5	Conveying course themes through interactive inquiry and didactic teaching	Individual	.34
		Average	.61
6	Holding the group learning environment	Individual	.27
		Average	.53

ICCs represent the average of rating 2 MBSR sessions per teacher. Individual ICC refers to ICC if ratings are done by a single evaluator. Average represents the ICC if ratings from 3 evaluators are averaged.

**Table 3.** MBI:TAC Audio Ratings Compared to Video Benchmark Ratings.

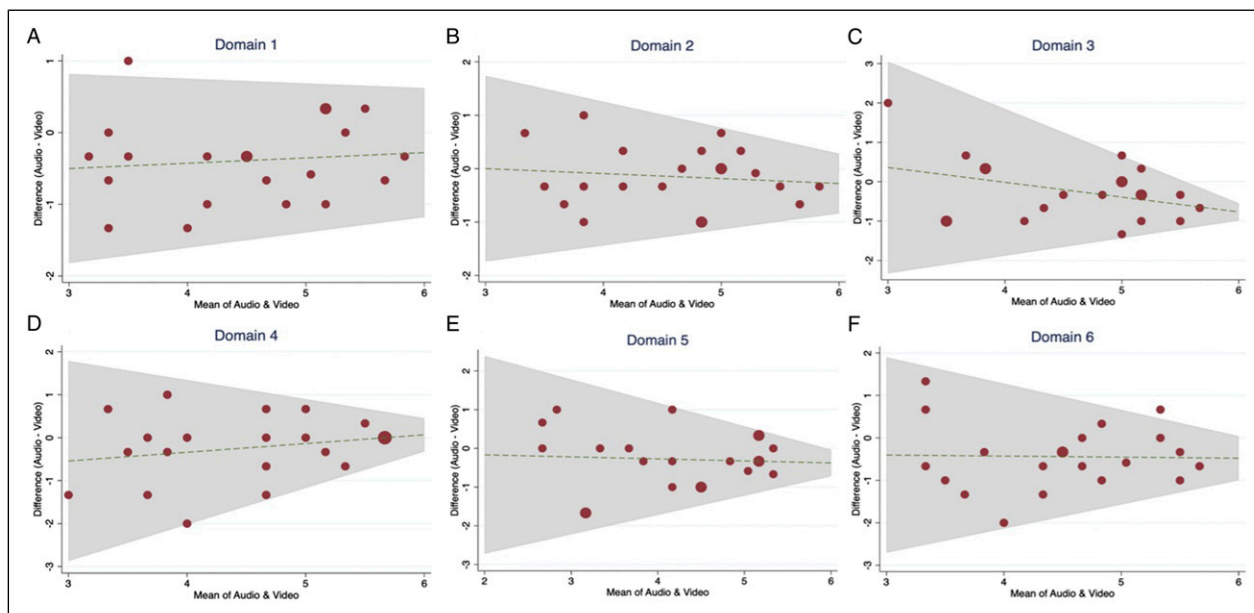
Domain	Mean Audio Rating (SD) [95% CI]	Mean Video Benchmark Ratings (SD) [95% CI]	Mean Difference Between Audio vs Video Benchmark Ratings (SD) [95% CI]	P-Value
1	4.32 (.93) [4.149, 4.483]	4.71 (.86) [4.551, 4.860]	.39 (.58) [.285, .494]	<.001
2	4.54 (.75) [4.399, 4.671]	4.69 (.78) [4.548, 4.830]	.15 (.52) [.061, .246]	.0014
3	4.50 (.72) [4.366, 4.626]	4.73 (.96) [4.558, 4.902]	.23 (.75) [.099, .369]	.0008
4	4.40 (1.00) [4.216, 4.577]	4.64 (.81) [4.490, 4.783]	.24 (.72) [.110, .370]	.0004
5	4.08 (.97) [3.909, 4.260]	4.38 (.98) [4.207, 4.559]	.30 (.70) [.172, .424]	<.001
6	4.31 (.82) [4.160, 4.456]	4.74 (.83) [4.590, 4.887]	.43 (.71) [.303-.558]	<.001

P-values are based on a paired t-test.

the bias was modest. The plots also suggested that the two rating methods tended to agree more closely for teachers with higher ratings.

We next looked at whether it was more challenging to rate less experienced teachers using the audio alone, compared with experienced teachers. In a linear mixed model with teacher years of formal practice as predictor and the difficulty of rating with audio-only recordings as the dependent variable, we found an association with years of formal meditation experience: for each additional 10 years of experience, the audio difficulty decreased by  $-.16$  on the scale, 95% CI:  $-.33$

to  $-.002$ ,  $P = .046$ , (Figure 2). For a similar model with years of teacher MBI experience as the predictor, audio rating tended to be perceived as easier with additional years of MBI teaching experience of the teacher being rated (for each additional 10 years of MBI experience, the audio rating difficulty decreased, with model coefficient:  $-.19$ , 95% CI:  $-.47$  to  $+.08$ ,  $P = .18$  (Figure 3), but this was not statistically significant. Considering instead the number of years of formal experience of the evaluators, there was some indication that the difficulty of using audio rather than video increased with more years of evaluator meditation experience, though the



**Figure 1.** Bland Altman Plots of Agreement Between MBI:TAC Ratings Using Audio Recordings and Video Recordings.

Each panel compares final audio and video ratings using Bland Altman plots for each domain of the MBI:TAC. Panel A represents MBI:TAC domain 1, panel B represents domain 2, panel C represents domain 3, panel D represents domain 4, panel E represents domain 5, and panel F represents domain 6. In each plot, the average final audio rating is compared with the video benchmark. The x-axis demonstrates the differences between the final audio vs video ratings, while the y-axis shows the mean of the audio and video ratings. Dots above 0 on the y-axis indicate the audio rating was higher than the video rating for the same teacher, whereas dots below 0 on the y-axis indicate the audio rating was lower than the video rating. The green dotted line indicates the linear relationship between the paired difference and paired average. The grey zone represents the limits of agreements adjusted from a regression model when a linear relationship between the paired difference and paired average exists. Where it occurs, narrowing of the grey zone with higher mean ratings indicates better agreement between audio and video ratings for teachers with higher ratings.

confidence interval was wide and it was not statistically significant: for each additional 10 years of mediation experience, the difficulty increased by .39, 95% CI:  $-.36$  to  $+1.15$ ,  $P = .31$ . (Figure 4). Similarly, for each additional 10 years of MBI teaching experience of the evaluator, the difficulty of using audio for ratings increased by .71, 95% CI:  $-.42$  to  $+1.81$ ,  $P = .22$  (Figure 5).

### Qualitative Analysis

In analyzing interviews with 8 evaluators to explore the experience of using audio-only recordings for rating using the MBI:TAC, we identified 3 themes: (1) video recordings were particularly helpful when rating less skillful teachers, (2) video recordings tended to provide a more complete picture for rating, and (3) audio rating had some positive features.

### Video Recordings Were Particularly Helpful When Rating Less Skillful Teachers

Many evaluators felt that rating less competent teachers using audio alone was more difficult than using video:

The second teacher [review] that I did, with audio alone, really was challenging and I didn't experience that teacher as an

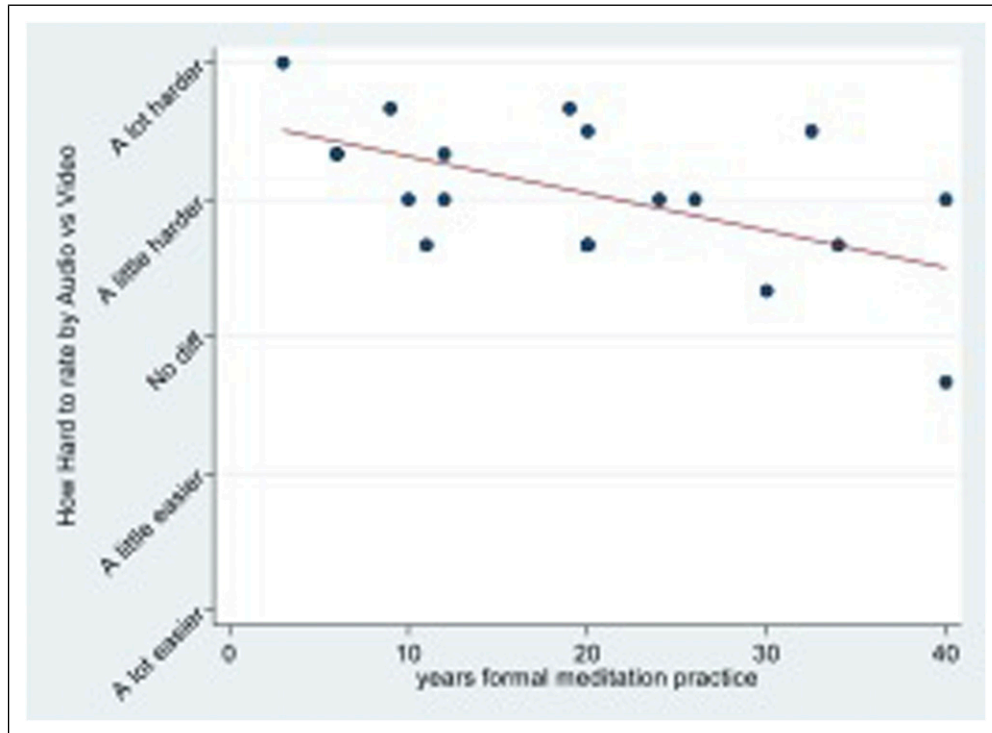
experienced teacher. I would have really liked to have seen them in action, because I feel like there's a lot of information available in the body that I didn't have access to. And just their language, it didn't sit well with me.... It was a challenging rating experience. (Interview 2, with female living in the United States with 11 years of mindfulness teaching experience)

Most evaluators felt that the visual component was less important for reviewing more advanced teachers, because they could measure the teachers' embodiment of mindfulness through the sound of their voices and get a sense of the teachers' "presence" through the audio recording. Likewise, interpersonal dynamics between the group and teacher could be noted via audio recordings, while visual information was less necessary to develop a clear sense of the interaction with advanced teachers.

### Video Recordings Tended to Provide a More Complete Picture for Rating

While evaluators had varying opinions regarding how significant visual data were during the MBI:TAC rating process, all 8 interviewees acknowledged that video added more sensory information than audio-only. Six out of 8 noted that completing the MBI:TAC ratings using the audio





**Figure 2.** Increased difficulty of rating with audio recording alone based on years of meditation practice of teacher being rated. Each dot represents 1 teacher who was rated, with the average across all evaluators of responses to this question: “For the teacher you just rated, what was your experience using audio for accurate MBI:TAC ratings compared to how you think it would be if you had a video recording?” using a response scale ranging from 1 = a lot easier to rate by audio rather than video to 5 = a lot harder to rate by audio rather than video. The x-axis shows the teacher’s years of formal meditation practice. The red line is a simple least squares fitted line. The *P*-value for the association between years of formal meditation practice and difficulty of rating by audio in a linear mixed model was  $P = .046$ .

format was more difficult than the video due to the lack of visual information. Some interviewees (3 out of 8) mentioned that to get the most accurate rating, video recordings should be used, since “everything is helpful” when optimizing accuracy (Interview 4, with female living in Spain with 7 years of mindfulness teaching experience). Another compared the visual information to additional pieces in a jigsaw puzzle:

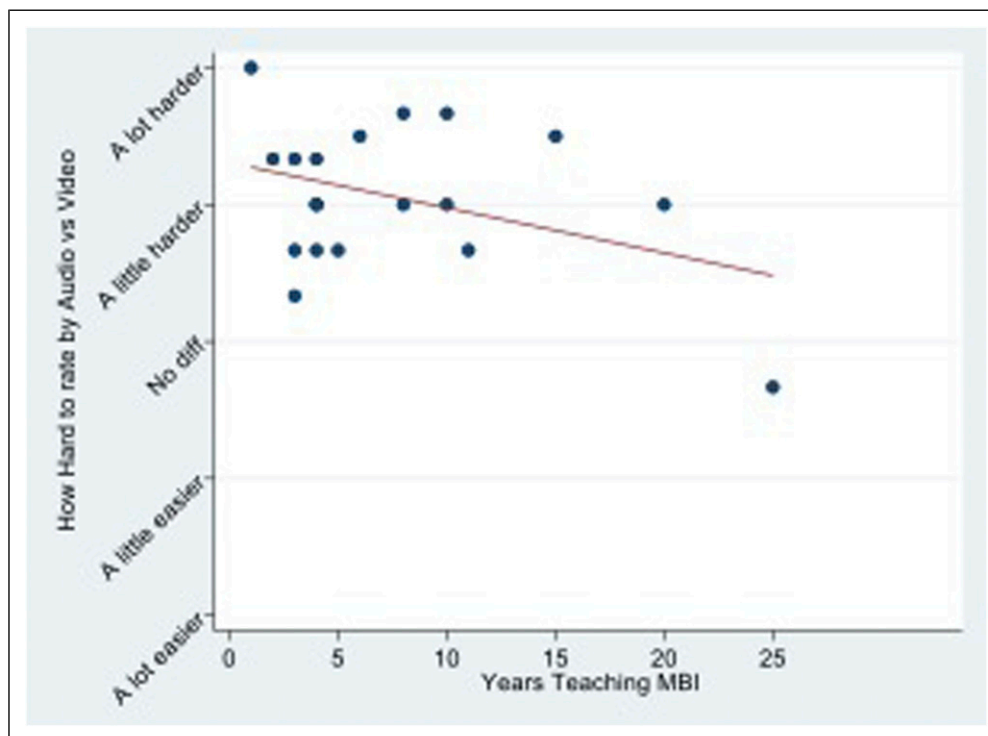
“I think it offers a complete picture, if you like. It’s like a jigsaw with many different parts, and then to get an overall sense, needing to see the detail of the pieces.” (Interview 6, with female living in the United Kingdom with 10 years of mindfulness teaching experience)

This same interviewee thought that the lack of video often left her questioning her final score:

I think there’s something in the fullness of being able to see and hear that helps to bring clarity as to which side of the line they may be on. With just the audio it was quite hard, because I felt like there was quite a lot of borderline.... It was like I needed more information to feel really sure [of] where I was placing people. (Interview 6)

As noted in the quantitative analysis, average audio ratings tended to be lower than video ratings. Without being aware of these data, this possibility was mentioned by some evaluators who hypothesized that they scored teachers lower when they lacked visual information:

“I might have graded higher if I could have seen the person and saw embodiment, for example, rather than just felt it.” (Interview 1, with female living in the United States with 22 years of mindfulness teaching experience)



**Figure 3.** Increased difficulty of rating with audio recording alone based on years of mindfulness-based intervention teaching experience of teacher being rated.

Each dot represents a teacher who was rated, with the average across all evaluators of responses to the question: “For the teacher you just rated, what was your experience using audio for accurate MBI:TAC ratings compared to how you think it would be if you had a video recording using a response scale from 1 = a lot easier to rate by audio rather than video, to 5 = a lot harder to rate by audio rather than video. The x-axis shows the teacher’s years of experience teaching MBI. The red line is a simple least squares fitted line. The P-value for the association between years of MBI teaching and difficulty of rating by audio in the linear mixed model was  $P = .18$ .

Some evaluators noted that the lack of visual information made the interpersonal relationships seem flat. Most evaluators described how the visual component created a more complete understanding of interpersonal relationships, class organization, visual displays, and the group mindfulness practices.

“There’s so much of communication that’s physical, not words, and you miss that whole piece. So, was that teacher leaning forward? Were they leaning back? Did their face look like they were interested? Did the laughter look like it was uncomfortable laughter or like it was natural?” (Interview 1).

### Audio Rating Had Some Positive Features

A few of the interviewees said that visual information was distracting in some cases or could bias or unnecessarily influence the rater: “How old somebody is, or their clothing, or whatever.... I think the video is more likely, for myself, to produce more snap judgments” (Interview 8, with male living in the United States with 12 years of mindfulness teaching experience).

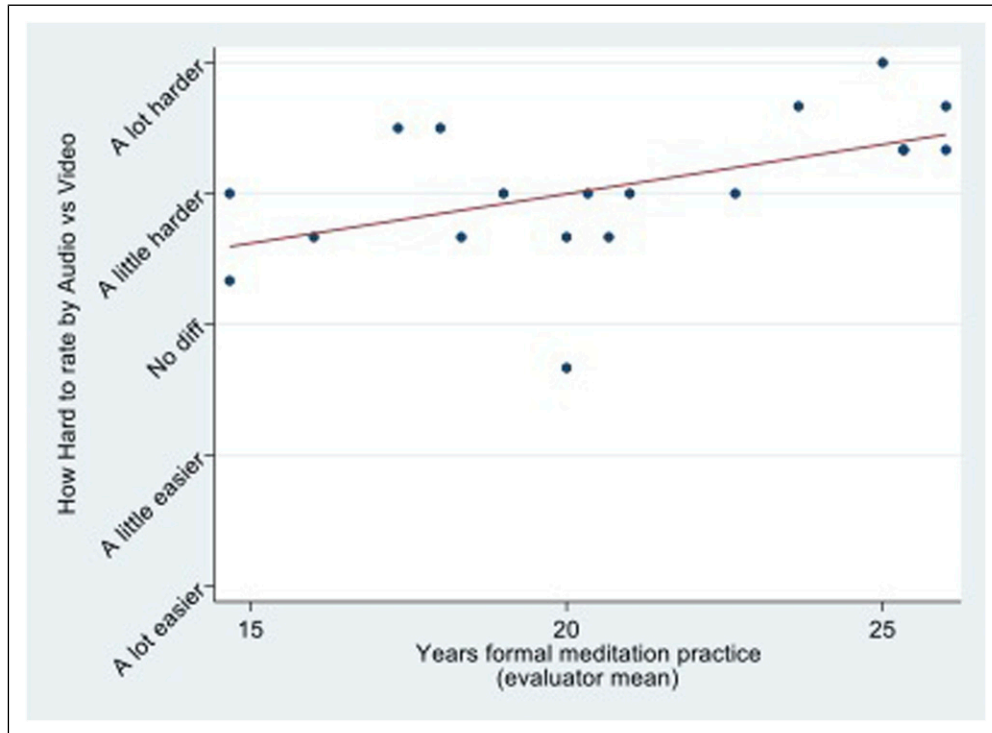
A couple of evaluators noted that increased MBI:TAC rating experience using a particular recording format was likely to be a more important factor in increasing accuracy than the specific recording format of recording used, although they still acknowledged that using video was easier in some cases.

A few interviewees noted other positive aspects of using the audio recordings instead of videos. For example, 1 explained that the audio recording may actually force the rater to be more present and really listen to what is being said.

### Additional Qualitative Data Findings

The 4 MBI: TAC domains that were most frequently mentioned as more difficult to rate via audio compared to video were ability to relate to the students (2), embodiment of mindfulness (3), inquiry (5), and holding the group environment (6). Evaluators who assessed MBI sessions in their second language felt that video format provided additional information for language comprehension, though they did not see this as an important barrier to using audio-only for ratings. One of these evaluators reported that although she was





**Figure 4.** Increased difficulty of rating with audio recording alone based on years of meditation practice of evaluator making rating. Each dot represents 1 teacher who was rated, with the average across all evaluators of responses to this question: “For the teacher you just rated, what was your experience using audio for accurate MBI:TAC ratings compared to how you think it would be if you had a video recording?” using a response scale ranging from 1 = a lot easier to rate by audio rather than video to 5 = a lot harder to rate by audio rather than video. The x-axis shows the evaluators’s years of formal meditation practice. The red line is a simple least squares fitted line. The *P*-value for the association between years of formal meditation practice and difficulty of rating by audio in a linear mixed model was  $P = .31$ .

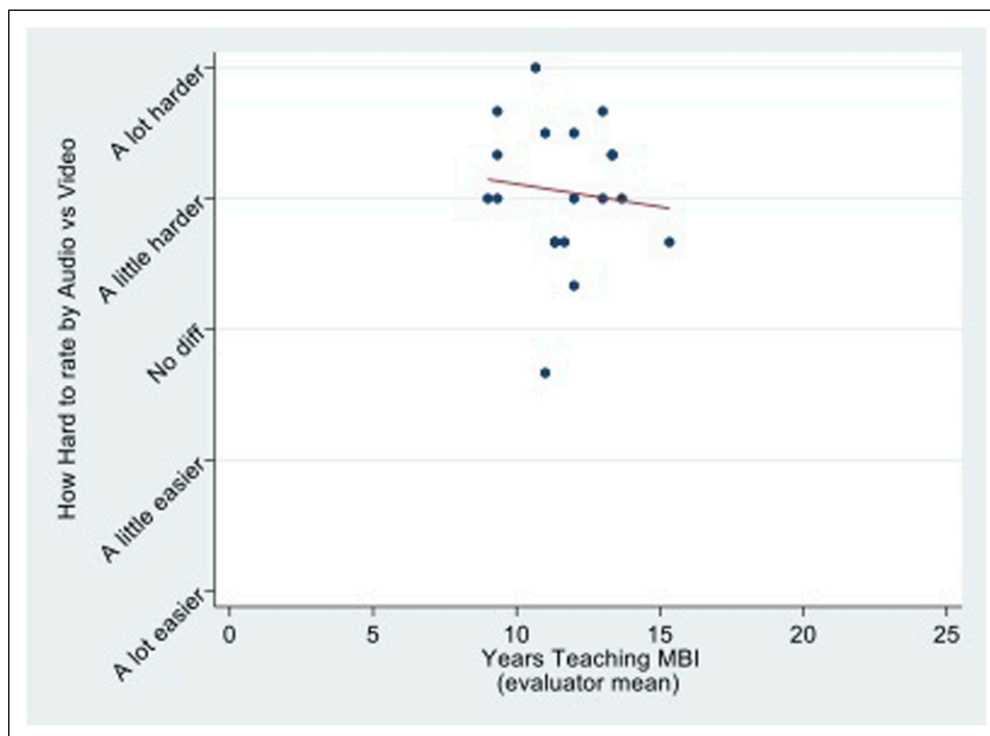
initially worried about the quality of audio-ratings given the language difference, she found that rating with audio was not as difficult as she had expected.

## Discussion

We found evidence that using a single evaluator with audio recordings to perform an MBI:TAC rating generally resulted in low ICCs. However, when a panel of 3 evaluators was used and ratings were averaged, ICCs were above .5, indicating relatively good inter-rater reliability. An analogy for the difference between individual and panel ratings is the way ice-skating performances are scored by trained judges. If the skater is scored by a single judge, the inter-rater reliability of the score is expected to be low. For this reason, a panel of judges is used, instead, and the ratings combined, providing better inter-rater reliability for the score of the performance. Our findings suggest that use of the MBI:TAC with audio recording is feasible, but averaging more than 1 rating is desirable for good inter-rater reliability. Although we did not directly assess the use of 2 evaluators per teacher, ICCs would be expected to be in-between these results.

Overall, ICCs of audio recordings were lower than ratings of video recordings. Ratings of the same teachers using video recordings had ICCs in the .6-.8 range using an average of multiple evaluators. While the differences in average scores on the MBI:TAC between ratings of video and audio recordings were modest, we found a fairly consistent trend toward lower ratings with audio recordings. This was consistent with the views expressed by some evaluators in interviews, several of whom had concerns that they might be scoring teachers lower without the additional information from the video recordings.

Bland-Altman plots provided evidence that ratings converged more closely for teachers who received higher scores on the MBI:TAC. We also found that teachers’ years of mindfulness practice was correlated with increased ease in rating their audio-recorded sessions. These quantitative findings were consistent with qualitative data from interviews, in which several evaluators reported that they felt the video information was particularly important when rating less experienced teachers. Taken together, these findings suggest that use of audio-only recordings for MBI:TAC ratings may be most appropriate when rating experienced teachers, for example, in the context of research studies. On the other hand,



**Figure 5.** Increased difficulty of rating with audio recording alone based on years of mindfulness-based intervention teaching experience of evaluator making rating.

Each dot represents a teacher who was rated, with the average across all evaluators of responses to the question: “For the teacher you just rated, what was your experience using audio for accurate MBI:TAC ratings compared to how you think it would be if you had a video recording using a response scale from 1 = a lot easier to rate by audio rather than video, to 5 = a lot harder to rate by audio rather than video. The x-axis shows the evaluator’s years of experience teaching MBI. The red line is a simple least squares fitted line. The  $P$ -value for the association between years of MBI teaching and difficulty of rating by audio in the linear mixed model was  $P = .22$ .

using audio recordings may be more problematic when rating teachers-in-training. The overall tendency for ratings from audio recordings to be slightly lower might be best considered in the context of how ratings from audio recordings might be used. For example, in teacher training this might mean adjusting feedback for what might be expected to be slightly lower scores when using audio recordings. When comparing ratings for teachers between research studies that used different recording media (video or audio), our findings provide some guide to adjustments that might be made to assess whether MBI:TAC scores were similar.

While the embodiment domain had the poorest ICC in this audio sub-study, it also had the lowest level of interrater agreement during the initial development of the MBI:TAC when evaluators compared rating MBI sessions using video recordings to live observation.<sup>9</sup> Crane et al.,<sup>9</sup> found that embodiment was the most challenging domain to articulate and the most open to interpretation. Our findings further support this original finding as the embodiment of mindfulness was the most difficult to rate reliably using the audio-only recording format. However, while the interviewees identified the domains which had the lowest ICCs, such as

the embodiment of mindfulness, their order of difficulty was not identical to the ICC findings. For example, the ICC associated with relational skills was much higher than the evaluators hypothesized in the qualitative interviews. This observation highlights the possibility that there may have been sufficient data in the audio recordings to evaluate interpersonal abilities, even if the evaluators found the process more difficult.

In other research that assessed the optimal means of recording medical group sessions for evaluation, there have been variable findings. Some studies have found that the process of rating such sessions is different for certain scales when sessions are recorded using the audio vs video format, while others have not.<sup>14-19</sup> Most studies exploring this topic found a non-significant difference in clinical ratings between audio-recorded and video-recorded clinical encounters.<sup>16-18</sup> One study even favored the audio-recorded sessions over video, noting that the visual information increased rating time and complexity when assessing communication between oncology patients and their physicians using the Cancode interaction system,<sup>15</sup> and that the intra-rater reliability scores were similar between recording formats.<sup>15</sup> However,

among these studies, a few aspects of the patient-provider relationship and communication, namely confrontation among empathic communication<sup>16</sup> and patronizing tone,<sup>18</sup> were rated differently depending on the recording format.

There were several limitations of this study. The MBSR teachers who were evaluated were predominantly rated within the upper 50% competency level. Our data is thus less informative about MBI:TAC assessments of MBI teachers with limited experience. Also, the number of teachers we evaluated was not large. Additional research may help to further define ICC values when using the MBI:TAC with audio recordings.

In summary, results from this pilot project suggest that audio recordings are adequate for research purposes in order to assess MBI teacher competency. Video recording appears to be optimal, when feasible, particularly when using the MBI:TAC for teacher training purposes.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: National Institutes of Health, National Center for Complementary and Integrative Health 5R34AT008948 (Hecht/Brewer), K24 AT007827 (Hecht), and T32 AT003997 (Hecht/Adler).

### ORCID iDs

Erin Floyd  <https://orcid.org/0000-0001-5432-528X>

Rebecca S. Crane  <https://orcid.org/0000-0003-3605-0256>

Willem Kuyken  <https://orcid.org/0000-0002-8596-5252>

Frederick M. Hecht  <https://orcid.org/0000-0002-5782-1171>

### References

- Grossman P, Niemann L, Schmidt S, Walach H. Mindfulness-based stress reduction and health benefits. *J Psychosom Res.* 2004;57(1):35-43.
- Cherkin DC, Sherman KJ, Balderson BH, et al. Effect of mindfulness-based stress reduction vs cognitive behavioral therapy or usual care on back pain and functional limitations in adults with chronic low back pain: A randomized clinical trial. *JAMA.* 2016;315(12):1240-1249.
- Khoury B, Sharma M, Rush SE, Fournier C. Mindfulness-based stress reduction for healthy individuals: A meta-analysis. *J Psychosom Res.* 2015;78(6):519-528.
- Kuyken W, Warren FC, Taylor RS, et al. Efficacy of mindfulness-based cognitive therapy in prevention of depressive relapse: An individual patient data meta-analysis from randomized trials. *JAMA Psychiatr.* 2016;73(6):565-574.
- National Institute for Health and Care Excellence (NICE). *Depression in adults: Treatment and management.* UK: NICE: 2022. <https://www.nice.org.uk/guidance/ng222/resources/depression-in-adults-treatment-and-management-pdf-66143832307909> (Accessed April 6, 2023).
- Moore GF, Audrey S, Barker M, et al. Process evaluation of complex interventions: Medical research council guidance. *BMJ.* 2015;350:h1258.
- Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: The new medical research council guidance. *BMJ.* 2008;337:a1655.
- Crane RS, Kuyken W, Williams JM, Hastings RP, Cooper L, Fennell MJ. Competence in teaching mindfulness-based courses: Concepts, development and assessment. *Mindfulness (N Y).* 2012;3(1):76-84.
- Crane RS, Eames C, Kuyken W, et al. Development and validation of the mindfulness-based interventions - teaching assessment criteria (MBI:TAC). *Assessment.* 2013;20(6):681-688.
- Crane RS, Bartley T, Evans A, Karunavira KV, Sansom S, Siverton S, Soulsby J, Williams V, Kuyken W, et al. *Mindfulness-based interventions teaching assessment criteria (MBI:TAC).* 3rd ed. Bangor, UK: Centre for Mindfulness Research and Practice, Bangor University; 2021. <https://mbitac.bangor.ac.uk/documents/MBITACmanual0517.pdf> (Accessed April 6, 2023).
- Crane RS, Hecht FM. Intervention integrity in mindfulness-based research. *Mindfulness (N Y).* 2018;9(5):1370-1380.
- Bland JMA, Douglas G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;327:307-310.
- Saldana J. *The Coding Manual for Qualitative Researchers.* 3rd ed. London, UK: Sage; 2016.
- Gionfriddo MR, Branda ME, Fernandez C, et al. Comparison of audio vs audio + video for the rating of shared decision making in oncology using the observer OPTION(5) instrument: An exploratory analysis. *BMC Health Serv Res.* 2018;18(1):522.
- Dent E, Brown R, Dowsett S, Tattersall M, Butow P. The Cancode interaction analysis system in the oncological setting: reliability and validity of video and audio tape coding. *Patient Educ Counsel.* 2005;56(1):35-44.
- Nicolai J, Demmel R, Farsch K. Effects of mode of presentation on ratings of empathic communication in medical interviews. *Patient Educ Counsel.* 2010;80(1):76-79.
- Weingarten MA, Yaphe J, Blumenthal D, Oren M, Margalit A. A comparison of videotape and audiotape assessment of patient-centredness in family physicians' consultations. *Patient Educ Counsel.* 2001;45:107-110.
- Williams K, Herman R, Bontempo D. Comparing audio and video data for rating communication. *West J Nurs Res.* 2013;35(8):1060-1073.
- Riddle DL, Albrecht TL, Coovert MD, et al. Differences in audiotaped vs videotaped physician-patient interactions. *J Nonverbal Behav.* 2002;26(4):219-239.