Journal of Nonverbal Behavior (2021) 45:53–66
https://doi.org/10.1007/s10919-020-00343-1

ORIGINAL PAPER

# Predictive Validity of Thin Slices of Verbal and Nonverbal Behaviors: Comparison of Slice Lengths and Rating Methodologies

Michael Z. Wang[1] · Katrina Chen[1] · Judith A. Hall[1]

Accepted: 24 September 2020 / Published online: 2 November 2020

**Abstract**
Thin slices, or excerpts of behavior, are commonly used by researchers to represent behaviors in their full stimulus. The present study asked how slices of different lengths and locations, as well as different measurement methodologies, influence correlations between the measured behavior and different variables (predictive validity). We collected self-rated, perceiver-rated, and objectively measured data on 60 participants who participated in a 5-min interaction with a confederate on video. These videos were split into five 1-min slices and rated for verbal and nonverbal behaviors via global impressions, using the same rater for all five slices and also using a different rater for each slice. For single slices, results indicated no clear pattern for optimal slice locations. In general, single slices had weaker predictive validity than the total. Slices of 2 or 3 min were, in general, equal to 5-min total in predictive validity. The magnitude of correlations was similar when same versus different coders were used, and the predictive validity correlations of the two methods covaried strongly across behavior-outcome variable combinations.

**Keywords** Thin slices · Nonverbal and verbal behavior · Predictive validity · Global impressions

## Introduction

Researchers have long known that social behaviors reliably correlate with various personality traits, abilities, and life outcomes (Ambady and Rosenthal 1992, 1993; Harker and Keltner 2001; Godfrey et al. 1986). For example, if an individual frequently smiles, laughs, or jokes around others, there is a good chance he or she is also extraverted, socially skilled, and has a great social life (Keltner and Bonanno 1997; Ruch and Deckers 1993; Yip and

✉ Michael Z. Wang
mikewang1990@gmail.com

✉ Judith A. Hall
j.hall@northeastern.edu

[1] Department of Psychology, Northeastern University, 125 NI, Boston, MA 02115, USA

Martin 2006). But how much of a behavior needs to be observed before it reliably correlates with a person's attributes?

Often, researchers use thin slices for behavioral measurement. A thin slice is an excerpt of behavior that is shorter than the total duration of the behavior the researcher has at hand, whether it consists of video, audio, or transcribed text (Ambady et al. 2000). Depending on the corpus at hand and the author's research purposes, a thin slice could vary substantially in duration—for example, 50 ms was used by Rule and Ambady (2008) while 10 min was used by Hirschmann et al. (2018). Thin slices have a long history in behavioral research (e.g., Scherer 1972), and have mainly been used for pragmatic reasons like alleviating the labor required of coders or raters. For example, Levine and Feldman (1997) chose 15-s slices out of a 10-min interaction for measuring negative and positive affect, and Goh et al. (2019, Study 1) used 30-s slices out of interactions that lasted up to 5 min to examine how target persons talked to a gay student versus a straight student. These are just two examples of the great many studies that have successfully used thin slices (e.g., Ambady and Gray 2002; Ambady and Rosenthal 1993; Fowler et al. 2009; Houser et al. 2007; Kraus and Keltner 2009; Oltmanns et al. 2004). The literature clearly demonstrates that thin slices can produce results of sufficient magnitude to be published, but such studies do not help researchers make more informed choices about slice length because different slice lengths are rarely compared.

The present article looks at how well a shorter video slice compares to a longer slice and to the total behavior duration, in terms of correlation with different (called here "outcome") variables. We use the term "predictive validity" to describe such correlations, as did Murphy et al. (2019), to indicate that the correlation is with separately measured attributes of the target people, not other slices measuring the same behavior.

In addition to asking about slice length and location in relation to predictive validity, the present study addressed another of the many methodological questions facing researchers (Blanch-Hartigan et al. 2018), that is, does rater methodology determine the predictive validity of slice ratings? A researcher who wishes to code multiple slices within the whole must decide whether different raters will rate each consecutive slice or the same raters will rate each slice instead.[1] Therefore, we compared two methodologies: (1) the same rater watches a 5-min video and makes a rating after every consecutive min, or (2) a different rater watches and rates each of the same five 1-min slices of the 5-min video. In the former case, the ratings within a video are not independent because they are rated in sequence by the same rater, and in the latter case the ratings are independent because, for a given target person, each slice is rated by a different rater. This question matters to thin slice researchers because the former method is much more economical (fewer raters are needed), yet it risks bias due to possible correlations between slices that reflect raters' memory biases from one slice to the next. On the other hand, using the same rater for all of the slices undoubtedly reduces noise, as that rater is self-calibrated whereas, in contrast, using a different rater for each slice introduces between-raters variance. (This is not an issue of how many raters should be assigned to a given slice, as that is a matter of interrater reliability, which is not the subject of the present article.)

[1] We use the term "rater" throughout because we used raters. Raters may be especially vulnerable to memory biases because ratings are arguably more subjective than coding methodologies such as counting occurrences or timing the duration of behavior. However, the general issues addressed in the present article are relevant to coders as well as raters.

In our study, 5-min videos were made of dyads having an unstructured conversation. The video was later coded for five verbal and nonverbal behaviors using these two methods. The individual slices, and the slices combined in cumulatively longer slices (e.g., slices 1 + 2), were then correlated with other variables in the database to assess the predictive validity of shorter and longer slices, and the location of slices, compared to the 5-min total.

## Comparative Thin Slice Research on Predictive Validity

Ambady and Rosenthal (1992) conducted a meta-analysis of correlations between what they called "expressive" behaviors coded from slices and a wide array of outcome variables. The authors found no association between slice length (which varied from under 30 s to 5 min) and the strength of the predictive correlations. Although this meta-analysis was important in showing that thin slices of measured behavior can predict other variables, it was not optimal for testing the impact of slice length—because the analysis was necessarily a comparison between (rather than within) studies, meaning that slice length was confounded with other study variables and therefore made for an imprecise test of the slice-length question.

Studies that varied slice length within the same study to compare the strength of prediction while keeping all else constant have yielded more conclusive insights. Ambady and Rosenthal (1993), in two studies of thin slices of teacher behavior predicting performance evaluations, compared 10-s, 5-s, and 2-s excerpts and found that although the correlations for longer slices were bigger, the longer slices did not predict to the criterion variable of teacher effectiveness significantly better than did the shorter slices. In Roter et al. (2011), three 1-min slices of verbal behavior (selected from early, middle, and late, as well as combined) were as predictive of independent judgments of rapport between clinicians and patients as was coding of the full 15-min interaction; also the single 1-min slices were not much different than the 3-min combined slice. Tskhay et al. (2017) obtained ratings of charisma from 5-s, 15-s, and 30-s silent slices from a 1-min video and found not much difference according to slice length in their correlations with independent ratings of leadership potential and with several other variables including gender, eye contact, wearing glasses, and physical attractiveness. Finally, Murphy et al. (2019) examined predictive validity in five studies for six nonverbal behaviors (nodding, smiling, gesturing, gazing, self-touch, and speaking time). While 1-min slices were somewhat worse in predicting a highly varied list of outcome variables than the whole 5-min videos were, 2-min slices were nearly as predictive as the 5-min totals.

Thus, the existing studies comparing slices of different lengths in terms of predictive validity present an optimistic picture for researchers who contemplate using thin slices for their behavioral measurement.

## The Current Study

Data came from a larger study that included more participants, behaviors, and outcome variables than reported here (Wang and Hall 2020). Decision rules for selecting participants, behaviors, and outcome variables were made a priori and are described in a later section. The current study used methodology that was both similar to, and different from, existing comparative thin slice studies.

### Behaviors Measured

Like previous studies (e.g., Murphy et al. 2019), we used 1-min slices and included some of the same behaviors previously used (smiling and nodding) while adding new ones (leaning in, humor/telling stories, and speaking about self).

### Outcome Variables

The outcome variables (i.e., variables that were not the rated nonverbal or verbal behaviors), were mostly different from those in previous such studies. We included self-reported and perceiver-rated variables, age, and an emotion recognition test.

### Rating Methodology

Methodology used in thin slice measurement can consist of frequency counts (e.g., number of nods measured by humans or machines), timing measured by humans or machines (e.g., gaze duration or time speaking), or global impressions on a rating scale (e.g., rating smiling from "not at all" to "very much"). The comparative thin slice literature has used all of these methods, but method choice has not been subjected to systematic analysis. The present study used global ratings, which is an ecologically valid methodology in that it approximates what people would do if they were evaluating someone's behavior in a real-life interaction (i.e., mentally combining their impression of duration, frequency, and intensity into one global impression).

As previewed above, we obtained slice ratings that varied in their independence from each other. In one of these approaches, the same two raters rated all five slices in order (independently from each other). In the second approach, a different rater rated each of the five slices according to a Latin Square design that ensured equal representation of each rater across targets and across slices.

## Method

### Overview

The study consisted of a laboratory session in which a 5-min interaction was video-recorded between a participant and a confederate. The participant provided data on a battery of instruments, and naïve perceivers made ratings of likeability and intelligence. Finally, trained raters analyzed the 1-min slices for verbal and nonverbal behaviors.

### Participants

The main study's participants were 152 students (40 male) at Northeastern University, either recruited from introductory psychology for partial course credit or via flyers on campus for monetary compensation (Wang and Hall 2020). For purposes of the present study, 60 of these participants were randomly selected. The average age was 20.32 years (SD = 2.67, range = 18–29); ethnicity was 46.7% Asian/Pacific Islander, 43.3% White/Caucasian, 6.7% Hispanic/Latin American, 3.3% Middle Eastern, 1.7% Black/African American, and 5.0% other ethnicity/race.

Other students from the same university served as 14 confederates (5 male; *M* age = 19.71), 40 video perceivers (17 male; *M* age = 19.50), and 7 behavior raters (1 male).

## Procedure

In the videotaped interactions with a confederate, both were instructed to talk about anything they liked. Participants filled out several surveys before and after the interaction.

## Materials for Participants

### Demographic Profile

Participants were asked their major, university year, date of birth, gender, and ethnicity/race.

### Social Life Quality Questionnaire

Eight self-rated items assessed participants' social lives (social life defined as "the time spent enjoying oneself with friends, acquaintances, and other people") in general, at work, with friends, and at the dormitory along two domains ("quality of social life" and "frequency of socializing"). Quality of social life was rated from 1 (*not very good*) to 5 (*very good*) and the frequency of socializing from 1 (*hardly ever*) to 5 (*all the time*). Items were averaged into one score and higher scores indicated greater quality of social life ($\alpha = .83$, $M = 3.93$, SD = .60).

### Geneva Emotion Recognition Test- Short

This 42-item test measured emotion recognition ability based on watching videotaped excerpts of target people expressing 14 emotions through face, body, and voice quality (Schlegel et al. 2014). A higher score indicated greater emotion recognition ability ($M = 26.32$, SD= 5.63).

### Narcissistic Personality Inventory-16

This scale (Ames et al. 2006) consisted of 16 pairs of items and participants had to choose which statement they identified with more in each pair (e.g., "everybody likes to hear my stories" or "sometimes I tell good stories") to assess self-reported levels of narcissistic personality. Items were scored on an answer key and correct answers were added up and divided by 16 to display a proportional score of narcissistic personality. A higher score indicated higher levels of narcissistic personality (*KR-20* = .76, $M = .28$, SD = .20).

### Interpersonal Reactivity Index (Perspective Taking Subscale)

This scale (Davis 1983) consisted of seven items rated on a 5-point rating scale from 1 (*does not describe me well*) to 5 (*describes me very well*) to assess the self-reported ability to adopt the point of view of others. Items were averaged (two items were reverse-coded) together to display a mean score of perspective taking. A higher average score indicated greater perspective taking ability ($\alpha = .73$, $M = 3.76$, SD = .61).

### Social Skills Inventory (Emotional Control Subscale)

This scale (Riggio 1986) consisted of 15 items self-rated on a 5-point scale from 1 (*not at all like me*) to 5 (*exactly like me*) to assess one's ability to control the display of emotions to others (e.g., hiding feelings of sadness from others). A higher score indicated greater control over the display of one's emotions ($\alpha = .79$, $M = 45.70$, SD = 9.14).

### Perceiver Ratings of Likeability and Intelligence

No perceivers were participants or confederates in the dyadic interactions. Two 30-s video slices were taken from each interaction video: One started at the 1-min mark and one started at the 3-min mark. Only the participant was in frame in the video recordings. Perceivers (who were naïve and untrained) were randomly assigned to rate all the interactions on one of these two slices, on either likeability or intelligence. The likeability questions were: "If you ever met this person in real life, how much would you want to see this person again?", "How likeable did you find this person to be?", and "If you ever met this person in real life, how likely would you guys be good friends?" Each was rated on a scale from − 10 (*not at all*) to 10 (*very much*) and the three items were averaged ($\alpha = .98$, $M = .78$, SD = 1.67). Intelligence questions were: "How intelligent did you find this person to be?", "How well-informed did you find this person to be?", and "How competent overall did you find this person to be?" Each was rated relative to other university students and the three were averaged ($\alpha = .96$, $M = 2.07$, SD = .98). For both likeability and intelligence, the two slices were averaged, further increasing reliability, so each participant had one perceiver likeability mean rating and one perceiver intelligence mean rating.

### Behavior Ratings

Behaviors were leaning in, nodding, smiling/laughing, humor/telling stories, and speaking about self. Each behavior was rated on a scale of 1 (*not at all*) to 5 (*constantly*).

### Same Rater for all Five Slices

Two research assistants rated behaviors from the videotapes. Neither participated in any other role in the study and were unacquainted with the participants. Each rated half of the interactions, so that all interactions were rated by one rater each. Raters rated their impressions after each 1-min slice, yielding five 1-min ratings for each behavior. Raters were instructed to form an impression one min at a time and to disregard impressions made in previous minutes. Raters watched a 1-min slice, rated leaning in and nodding (round 1 of rating), re-watched the same 1-min slice and rated humor/telling stories and smiling/laughing (round 2 of rating), and re-watched the same 1-min slice a third time and rated speaking about self (round 3 of rating). This pattern of watching the 1-min slice and rating behaviors in three rounds was conducted for all five 1-min slices (always in this order of behaviors to rate). Reliability was assessed by having the raters independently rate the same behaviors from 20 of the same videos. These ratings were then split by min slice and behavior and correlated between the two raters. The resulting correlations were then Fisher-z transformed for normalization, averaged, and these averages were turned back into correlations. We used these final correlations as indicators of reliability: leaning in ($r$ = .85), nodding ($r$ = .52), smiling/laughing ($r$ = .44), humor/telling stories ($r$ = .39), and speaking about self ($r$ = .68).

### Different Raters for Each of the Five Slices

Five research assistants rated the same behaviors according to a Latin Square design. Each of the five raters watched a different 1-min slice for each interaction, with the assignment scheme guaranteeing equal representation of raters across interactions and slices. Raters rated the same 1-min slice in three rounds for each slice: round 1 consisted of leaning in and nodding, round 2 consisted of humor/telling stories and smiling/laughing, and round 3 consisted of speaking about self. Reliability was assessed by having the raters independently rate the same behaviors for each min (one after the other) for all five min for 10 videos. Intraclass correlations between the five raters were calculated after being split by min slice and behavior. These intraclass correlations were then Fisher-z transformed, averaged, and these averages were turned back into correlations. We used these final correlations as indicators of reliability: leaning in ($r$ = .47), nodding ($r$ = .45), smiling/laughing ($r$ = .70), humor/telling stories ($r$ = .53), and speaking about self ($r$ = .67).

## Results

An a priori criterion was used for deciding which outcome variables from all of those measured in the main study would be included. As an operational criterion an outcome variable was included if the total (5-min) behavior was correlated with it at $p \leq .10$, two-tail, for either the same-rater (nonindependent) or the different-raters (independent) method.[2] Fifteen behavior-outcome variable combinations met this criterion and are shown in Table 1. The table shows the predictive validity correlations for individual slices, cumulative slices, and 5-min total. The table also shows the correlations for both same-rater and different-rater approaches. In addition, we present the median (absolute) correlation for each slice length and total, in order to better visualize the overall trends across slices and between rating approaches.

First, the predictive validity of single slices was generally somewhat lower than that of the total as summarized in the medians at the bottom of the table. Second, there was no consistent tendency for any given 1-min slice to be better than others for predictive validity. Variation among the single-slice correlations within each coding approach did not fit any particular trend, as also indicated by the median correlations at the bottom of the table. There was some variation, however, according to specific combinations of behaviors and outcome variables. For example, for speaking about self predicting perceiver-rated intelligence, there was a clear tendency for later slices to be more predictive than earlier slices.

Third, whereas individual slices were generally somewhat weaker than the total in predicting the outcome variables, the cumulative slices did a better job. For same-rater

[2] In the main study, other behaviors and potential outcome variables were measured, but they are not included here because they did not meet this criterion.

**Table 1** Predictive validity of individual and cumulative slices

| Variables | Individual slices | | | | | Cumulative slices | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | Slice 1 | Slice 2 | Slice 3 | Slice 4 | Slice 5 | Slice 1 + 2 | Slice 1 + 2 + 3 | Slice 1 + 2 + 3 + 4 | |
| Leaning in predicting emotion recognition test | | | | | | | | | |
| Same rater | − .18 | − .37** | − .28* | − .28* | − .22 | − .30* | − .31* | − .31* | − .30* |
| Different raters | − .17 | − .32* | − .24$^{\dagger}$ | − .21 | − .35** | − .28* | − .30* | − .30* | − .34** |
| Leaning in predicting age | | | | | | | | | |
| Same rater | .20 | .24$^{\dagger}$ | .19 | .14 | .23$^{\dagger}$ | .24$^{\dagger}$ | .24$^{\dagger}$ | .21 | .22$^{\dagger}$ |
| Different raters | .08 | .24$^{\dagger}$ | .11 | .11 | .26* | .19 | .18 | .17 | .21 |
| Nodding predicting perceiver-rated likeability | | | | | | | | | |
| Same rater | .18 | .31* | .31* | .20 | .16 | .28* | .31* | .30* | .28* |
| Different raters | .14 | .16 | .22$^{\dagger}$ | .15 | .11 | .18 | .23$^{\dagger}$ | .23$^{\dagger}$ | .22$^{\dagger}$ |
| Nodding predicting perceiver-rated intelligence | | | | | | | | | |
| Same rater | .16 | .16 | .28* | .24$^{\dagger}$ | .08 | .18 | .23$^{\dagger}$ | .25$^{\dagger}$ | .22* |
| Different raters | .22$^{\dagger}$ | .02 | .23$^{\dagger}$ | .06 | − .02 | .14 | .21 | .18 | .15 |
| Humor/telling stories predicting emotion recognition test | | | | | | | | | |
| Same rater | .26* | .02 | .18 | .23$^{\dagger}$ | .02 | .18 | .22$^{\dagger}$ | .26* | .24$^{\dagger}$ |
| Different raters | .18 | − .15 | .12 | .10 | − .06 | .02 | .08 | .11 | .08 |
| Humor/telling stories predicting narcissism | | | | | | | | | |
| Same rater | .24$^{\dagger}$ | .17 | .00 | − .05 | .13 | .25$^{\dagger}$ | .17 | .13 | .15 |
| Different raters | .02 | .16 | .08 | .13 | .33** | .12 | .13 | .17 | .27* |
| Humor/telling stories predicting perceiver-rated likeability | | | | | | | | | |
| Same rater | .32* | .33** | .30* | .34** | .44*** | .39** | .42*** | .47*** | .55*** |
| Different raters | .29* | .25$^{\dagger}$ | .20 | .39** | .10 | .36** | .38** | .48*** | .46*** |
| Humor/telling stories predicting perceiver-rated intelligence | | | | | | | | | |
| Same rater | .18 | .10 | .08 | .09 | .23$^{\dagger}$ | .16 | .15 | .17 | .22$^{\dagger}$ |
| Different raters | .34** | .08 | .06 | .43*** | .16 | .29* | .25$^{\dagger}$ | .40*** | .41*** |

**Table 1** (continued)

| Variables | Individual slices | | | | | Cumulative slices | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | Slice 1 | Slice 2 | Slice 3 | Slice 4 | Slice 5 | Slice 1 + 2 | Slice 1 + 2 + 3 | Slice 1 + 2 + 3 + 4 | |
| Smiling/laughing predicting perceiver-rated likeability | | | | | | | | | |
| Same rater | .34* | .45*** | .32* | .43*** | .32* | .43*** | .43*** | .45*** | .45*** |
| Different raters | .47*** | .48*** | .32* | .30* | .28* | .53*** | .53*** | .52*** | .52*** |
| Smiling/laughing predicting emotional control | | | | | | | | | |
| Same rater | − .24$^{\dagger}$ | − .20 | − .11 | − .24$^{\dagger}$ | − .10 | − .24$^{\dagger}$ | − .21 | − .23$^{\dagger}$ | − .22$^{\dagger}$ |
| Different raters | − .10 | − .16 | .03 | − .06 | − .12 | − .15 | − .10 | − .10 | − .12 |
| Smiling/laughing predicting perspective taking | | | | | | | | | |
| Same rater | .19 | .22$^{\dagger}$ | .19 | .20 | .20 | .22$^{\dagger}$ | .23$^{\dagger}$ | .23$^{\dagger}$ | .24$^{\dagger}$ |
| Different raters | .21 | .28* | .08 | .26* | .12 | .27* | .24$^{\dagger}$ | .27* | .26* |
| Smiling/laughing predicting social life quality | | | | | | | | | |
| Same rater | .14 | .24$^{\dagger}$ | .24$^{\dagger}$ | .30* | .17 | .21 | .24$^{\dagger}$ | .27* | .26* |
| Different raters | .10 | .23$^{\dagger}$ | .26* | .10 | − .01 | .18 | .24$^{\dagger}$ | .23$^{\dagger}$ | .19 |
| Speaking about self predicting perceiver-rated intelligence | | | | | | | | | |
| Same rater | .09 | .17 | .00 | .25* | .31 * | .16 | .11 | .18 | .24$^{\dagger}$ |
| Different raters | .07 | .13 | .20 | .30* | .36* | .13 | .21 | .29* | .38** |
| Speaking about self predicting age | | | | | | | | | |
| Same rater | .02 | .09 | .16 | .28* | .13 | .07 | .12 | .20 | .21 |
| Different raters | .06 | .17 | .32* | .14 | .13 | .16 | .28* | .28* | .28* |
| Speaking about self predicting narcissism | | | | | | | | | |
| Same rater | .00 | .10 | .16 | .27* | .07 | .07 | .12 | .19 | .19 |
| Different raters | .13 | .12 | .31 * | .00 | .15 | .16 | .28* | .22$^{\dagger}$ | .23$^{\dagger}$ |
| Median (absolute values) | | | | | | | | | |
| Same rater | .18 | .20 | .19 | .24 | .17 | .22 | .23 | .23 | .24 |
| Different raters | .14 | .16 | .20 | .14 | .13 | .18 | .24 | .23 | .26 |

**Table 1** (continued)

$N = 60$. Same rater means that a single rater coded all five slices of a given target, consecutively. Different raters means that each of a target's five slices was coded by a different rater. See text for methodology. The median pertains to the absolute values of the correlations in each column, calculated separately for same and different rater correlations

$^{\dagger}p < .10$ $^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

**Table 2** Vector correlations between correlations based on nonindependent versus independent ratings

| Slices | Vector correlation |
|---|---|
| Slice 1 | .79 |
| Slice 2 | .93 |
| Slice 3 | .79 |
| Slice 4 | .61 |
| Slice 5 | .76 |
| Cumulative slices 1 + 2 | .89 |
| Cumulative slices 1 + 2 + 3 | .89 |
| Cumulative slices 1 + 2 + 3 + 4 | .90 |
| Cumulative slices 1 + 2 + 3 + 4 + 5 (total) | .90 |

*N* = 15 (the number of pairs of behavior-outcome correlations in the vectors)

correlations, slices 1 + 2 on average predicted as well as total, while for different-raters correlations, slices 1 + 2 + 3 predicted as well as total. There was variation, again, for specific behavior-outcome variable combinations.

Fourth, correlations for same rater were not noticeably different from those based on different raters, as seen in the median correlations. Also, the patterning of the results for the two rating methods was very similar. As a way to quantify the similarity in pattern between the two rating methodologies, we calculated a vector correlation (cf. Back et al. 2008; Gosling et al. 2002) for each slice length, which depicts how the pattern of same-rater correlations matches that of the different-raters correlations, across the 15 combinations of behaviors and outcome variables in Table 1. These are shown in Table 2. The vector correlations were substantial, indicating that two rating methods produced predictive validity correlations that strongly covaried.

## Discussion

The goals of this study were to discover how much predictive validity is lost, if any, when using a thin slice as opposed to using the entire duration of recorded video material, and to compare two rating approaches in terms of their impact on predictive validity. Predictive validity was defined as the correlation of a measured behavior with a different variable, which in our case included self-reported social life quality, emotional control, narcissism, perspective taking, and age, as well as perceiver ratings of likeability and intelligence, and an emotion recognition test.

Trained raters used two different rating methodologies enabling us to examine a potentially important methodological issue: whether predictive validity is influenced by having the same rater versus different raters watch the slices within each interaction. There was not much difference in overall magnitude of prediction between these two methods, and furthermore the pattern of correlations between the two approaches was very similar. In choosing between these approaches, future researchers might prefer the same-rater (nonindependent) method, considering it is more labor- and time-efficient. Whether one method is more valid than the other is not clear, however. Using one rater may introduce some carryover from one slice to the next, but it also reduces random error introduced by having a different rater watch each slice. Of course, there remain other approaches yet to be compared, such as having a single rater do all of the first slices, then go back and do all of the second slices, and so forth, thus reducing memory bias while requiring fewer raters.

In terms of evaluating single and cumulative slices, we found that single slices predicted to the outcomes somewhat worse than the 5-min total did, and that there was no general pattern of superiority for specific slice locations, as the magnitude of correlations varied with both behaviors and outcome variables. Cumulative slices showed that, on average, either slices 1 + 2 or 1 + 2 + 3 predicted nearly as well as total, depending on method and, again, there was some variation depending on the behavior and the outcome variable. The overall conclusion, however, is that predictive validity can be retained with slice lengths shorter than the total.

### Limitations and Future Directions

Only a limited number of behaviors and outcome variables can be selected and examined within one study. In the present study, there was no bias in this selection because the study from which the data came was conducted for unrelated purposes and we used all of the variables from the study that fit a priori criteria. Nevertheless, limitations on generalization result from the specific combinations of behaviors, outcome variables, and slice lengths we examined.

Behaviors were measured in terms of holistic impressions. The use of impression ratings that require raters to integrate frequency, duration, and intensity into a global assessment has a well-established history (e.g., Ambady and Rosenthal 1993; Blanch-Hartigan et al. 2018; Briton and Hall 1995; Funder and Colvin 1991). While the present findings may help researchers who regularly employ global impressions of behavior, future thin slice validity studies may also want to count the frequencies or time behavior durations and compare.

Overall, this research was intended to help future researchers make decisions when using thin slices, hopefully resulting in more efficient and cost-effective research. In a single study, it is impossible to capture all the behaviors a researcher might wish to measure, all the other variables they might wish to correlate those behaviors with, all the durations of thin slices that researchers could select (5 s, 30 s, 1 min, etc.), nor all the possible durations of the "total" stimuli that different studies might have collected (5 min, 15 min, 45 min, etc.). Therefore, any one study will inevitably fall far short of answering all of the questions researchers have about the comparative utility of using thin slices. Because of this, it is important that repeated efforts are made that will, collectively, increase the range of content as well as replicability of thin slice validity investigations. As this small but growing literature expands, our hope is that researchers who rely on thin slices will be well-served by this repository of data in making logistical and data-driven thin slice decisions.

## References

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology, 32,* 201–271.

Ambady, N., & Gray, H. M. (2002). On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of Personality and Social Psychology, 83,* 947–961.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111,* 256–274.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64,* 431–441.

Ames, D. R., Rose, P., & Anderson, C. P. (2006). The NPI-16 as a short measure of narcissism. *Journal of Research in Personality, 40,* 440–450.

Back, M. D., Schmukle, S. C., & Egloff, B. (2008). How extraverted is honey.bunny77@hotmail.de? Inferring personality from e-mail addresses. *Journal of Research in Personality, 42,* 1116–1122.

Blanch-Hartigan, D., Ruben, M. A., Hall, J. A., & Schmid Mast, M. (2018). Measuring nonverbal behavior in clinical interactions: A pragmatic guide. *Patient Education and Counseling, 101,* 2209–2218.

Briton, N. J., & Hall, J. A. (1995). Gender-based expectancies and observer judgments of smiling. *Journal of Nonverbal Behavior, 19,* 49–65.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44,* 113.

Fowler, K. A., Lilienfeld, S. O., & Patrick, C. J. (2009). Detecting psychopathy from thin slices of behavior. *Psychological Assessment, 21,* 68–78.

Funder, D. C., & Colvin, C. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology, 60,* 773–794.

Godfrey, D. K., Jones, E. E., & Lord, C. G. (1986). Self-promotion is not ingratiating. *Journal of Personality and Social Psychology, 50,* 106–115.

Goh, J. X., Ruben, M. A., & Hall, J. A. (2019). When social perception goes wrong: Judging targets' behavior toward gay versus straight people. *Basic and Applied Social Psychology, 41,* 63–71.

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology, 82,* 379–398.

Harker, L., & Keltner, D. (2001). Expressions of positive emotion in women's college yearbook pictures and their relationship to personality and life outcomes across adulthood. *Journal of Personality and Social Psychology, 80,* 112–124.

Hirschmann, N., Kastner-Koller, U., Deimann, P., Schmelzer, M., & Pietschnig, J. (2018). Reliable and valid coding of thin slices of video footage: Applicability to the assessment of mother-child interactions. *Journal of Psychopathology and Behavioral Assessment, 40,* 249–258.

Houser, M. L., Horan, S. M., & Furler, L. A. (2007). Predicting relational outcomes: An investigation of thin slice judgments in speed dating. *Human Communication, 10,* 69–81.

Keltner, D., & Bonanno, G. A. (1997). A study of laughter and dissociation: distinct correlates of laughter and smiling during bereavement. *Journal of Personality and Social Psychology, 73,* 687–702.

Kraus, M. W., & Keltner, D. (2009). Signs of socioeconomic status: A thin-slicing approach. *Psychological Science, 20,* 99–106.

Levine, S. P., & Feldman, R. S. (1997). Self-presentational goals, self-monitoring, and nonverbal behavior. *Basic and Applied Social Psychology, 19,* 505–518.

Murphy, N. A., Hall, J. A., Ruben, M. A., Frauendorfer, D., Schmid Mast, M., Johnson, K. E., et al. (2019). Predictive validity of thin-slice nonverbal behavior from social interactions. *Personality and Social Psychology Bulletin, 45,* 983–993.

Oltmanns, T. F., Friedman, J. N., Fiedler, E. R., & Turkheimer, E. (2004). Perceptions of people with personality disorders based on thin slices of behavior. *Journal of Research in Personality, 38,* 216–229.

Riggio, R. E. (1986). Assessment of basic social skills. *Journal of Personality and Social Psychology, 51,* 649–660.

Roter, D. L., Hall, J. A., Blanch-Hartigan, D., Larson, S., & Frankel, R. M. (2011). Slicing it thin: New methods for brief sampling analysis using RIAS-coded medical dialogue. *Patient Education and Counseling, 82,* 410–419.

Ruch, W., & Deckers, L. (1993). Do extraverts "like to laugh"? An analysis of the Situational Humor Response Questionnaire (SHRQ). *European Journal of Personality, 7,* 211–220.

Rule, N. O., & Ambady, N. (2008). Brief exposures: Male sexual orientation is accurately perceived at 50ms. *Journal of Experimental Social Psychology, 44,* 1100–1105.

Scherer, K. R. (1972). Judging personality from voice: A cross-cultural approach to an old issue in interpersonal perception. *Journal of Personality, 40,* 191–210.

Schlegel, K., Grandjean, D., & Scherer, K. R. (2014). Introducing the geneva emotion recognition test: An example of Rasch-based test development. *Psychological Assessment, 26,* 666–672.

Tskhay, K. O., Zhu, R., & Rule, N. O. (2017). Perceptions of charisma from thin slices of behavior predict leadership prototypicality judgments. *The Leadership Quarterly, 28,* 555–562.

Wang, M. Z., & Hall, J. A. (2020). From lab to life: Impression management effectiveness and behaviors. *Social Influence, 15,* 46–63.

Yip, J. A., & Martin, R. A. (2006). Sense of humor, emotional intelligence, and social competence. *Journal of Research in Personality, 40,* 1202–1208.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.