Reliability and validity of nonverbal thin slices in social interactions

Article *in* Personality and Social Psychology Bulletin · February 2015

Impact Factor: 2.52 · DOI: 10.1177/0146167214559902 · Source: PubMed

8 authors, including:

Marianne Schmid Mast
University of Lausanne
116 PUBLICATIONS **1,538** CITATIONS
SEE PROFILE

Denise Frauendorfer
Université de Neuchâtel
11 PUBLICATIONS **99** CITATIONS
SEE PROFILE

Debra Roter
Johns Hopkins Bloomberg School of Public …
310 PUBLICATIONS **16,980** CITATIONS
SEE PROFILE

Laurent Son Nguyen
Idiap Research Institute & EPFL
15 PUBLICATIONS **51** CITATIONS
SEE PROFILE

Reliability and Validity of Nonverbal Thin Slices in Social Interactions

Personality and Social Psychology Bulletin
2015, Vol. 41(2) 199–213

DOI: 10.1177/0146167214559902
pspb.sagepub.com

Nora A. Murphy[1], Judith A. Hall[2], Marianne Schmid Mast[3], Mollie A. Ruben[2], Denise Frauendorfer[3], Danielle Blanch-Hartigan[4], Debra L. Roter[5], and Laurent Nguyen[6]

Abstract

Four studies investigated the reliability and validity of thin slices of nonverbal behavior from social interactions including (a) how well individual slices of a given behavior predict other slices in the same interaction; (b) how well a slice of a given behavior represents the entirety of that behavior within an interaction; (c) how long a slice is necessary to sufficiently represent the entirety of a behavior within an interaction; (d) which slices best capture the entirety of behavior, across different behaviors; and (e) which behaviors (of six measured behaviors) are best captured by slices. Notable findings included strong reliability and validity for thin slices of gaze and nods, and that a 1.5-min slice from the start of an interaction may adequately represent some behaviors. Results provide useful information to researchers making decisions about slice measurement of behavior.

Keywords

nonverbal behavior, coding, thin slices, reliability, validity

Received October 23, 2013; revision accepted October 18, 2014

More and more, it is shown that perceivers can glean accurate information about target persons' states, traits, and personal characteristics from very short excerpts—thin slices—of their behavior or appearance (Ambady, Bernieri, & Richeson, 2000; Ambady & Rosenthal, 1992). A host of studies indicate that, in some domains, shorter and longer slices produce inferences of similar accuracy. Ambady and Rosenthal demonstrated that accuracy rates of predicting various social and clinical outcomes did not significantly differ between 30-s and 5-min observations of expressive behavior. There was not much difference in accuracy between exposures of 5 min and exposures of 1 min in judging personality traits, and for some traits there was not much difference, if any, when exposures were reduced to 5 s (Carney, Colvin, & Hall, 2007). The accuracy of judging affect can be substantially above chance with audiovisual clips or photographs lasting only 2 s (Nowicki & Duke, 1994; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979) or even less than 1 s (Matsumoto et al., 2000). Accuracy in perception of sexual orientation and extraversion is above chance at 50 ms (Borkenau, Brecke, Möttig, & Paelecke, 2009; Rule & Ambady, 2008). In a quantitative review of 30 studies, there was not much difference in accuracy of judging prototypical facial expressions of emotion between exposures of less than 1 s and considerably longer exposures (Hall, Andrzejewski, Murphy, Schmid Mast, & Feinstein, 2008). Though accuracy may not always be very far above chance, especially at extremes of stimulus brevity, nevertheless it is clear that thin slices do carry valid information.

There is, of course, an extensive tradition of researchers doing behavioral coding on short excerpts of behavior (e.g., Ambady et al., 2000; Ambady & Rosenthal, 1992, 1993; Grahe & Bernieri, 1999; Hall, Roter, & Rand, 1981; Milmoe, Rosenthal, Blane, Chafetz, & Wolf, 1967; Murphy, 2007; Newton, Haviland, & Contrada, 1996; Tucker & Anders, 1998). The fact that such studies often produce significant correlations between the measured behaviors and other variables supports the validity of the thin-slice approach (Ambady et al., 2000; Roter, Hall, Blanch-Hartigan, Larson, & Frankel, 2011). And many studies using a lens model approach demonstrate that, indeed, behavior measured in thin slices can correlate significantly and often strongly with a construct of interest, such as personality (e.g., loud voice

[1]Loyola Marymount University, Los Angeles, CA, USA
[2]Northeastern University, Boston, MA, USA
[3]University of Lausanne, Switzerland
[4]National Cancer Institute, Bethesda, MD, USA
[5]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[6]Swiss Federal Institute of Technology in Lausanne (EPFL) and Idiap Research Institute, Lausanne, Switzerland

Corresponding Author:
Nora A. Murphy, Department of Psychology, Loyola Marymount University, 1 LMU Drive, Los Angeles, CA 90045, USA.
Email: nora.murphy@lmu.edu

correlates with extraversion, and eye gaze correlates with IQ (Borkenau, Mauer, Riemann, Spinath, & Angleitner, 2004; Murphy, 2007).

Thin Slices Predicting Within-Interaction Behavior

Inherent in the thin-slice approach is the question of whether the thin slice itself contains valid information about the behavior it is intended to capture. If thin slices can produce accurate inferences about affect, personality, and personal attributes, then perhaps thin slices can also produce accurate measurements of the behavior itself. In other words, is the measurement of a given behavior from a thin slice a valid substitute for the same behavior measured over a longer length of time? This question has rarely been addressed empirically, even though it comes up inevitably when researchers design the coding phase of their studies. If the thin slice can substitute for the whole, it would be an attractive conclusion for researchers for a number of reasons. On a practical level, behavioral coding is extremely laborious and time-consuming, as many researchers note (e.g., Gosling, John, Craik, & Robins, 1998; Murphy, 2005). For example, in one recent study, researchers coded over 7.9 million video frames of behavioral data for analysis of 138 participants in 36-min interactions (though no mention was made about the length of time to complete the coding; Fairbairn, Sayette, Levine, Cohn, & Creswell, 2013). Thus, using thin slices is an appealing option to nonverbal researchers as thin-slice behavioral coding reduces the time necessary to effectively measure behavior. Also, when participants serve as naïve raters of target interactions (as in Hall, Carter, & Horgan, 2001, Study 2), thin slices offer the advantage of efficiency as participants can view more targets in the same amount of time, reducing the number of participants required. Finally, authors are often challenged to justify the use of thin slices to represent longer behavioral streams, and investigating whether substituting thin slices for coding lengthier interactions might provide acceptable evidence.

Beyond practical applications, the study of thin-slice reliability and validity potentially relates to larger questions of behavioral consistency and personality. The person–situation debate as to whether traits predict behavior has a long history in both social and personality psychology (Kenrick & Funder, 1988; Leary & Hoyle, 2009; Mischel, 1968). Researchers investigating behavior–personality links emphasized how the aggregation of measurement, across situations, across participants, and across judges, would increase the reliability and generalizability of findings (Epstein, 1979, 1983a; Moskowitz & Schwarz, 1982). Today, most researchers (though not all) would agree that there is some stability to personality traits (Funder, 2006, 2009; Leary & Hoyle, 2009; Tellegen, 1991). The analogy between behavioral consistency in personality to the present work is that rather than studying individuals and variability or consistency in personality, we investigated the variability or consistency of behavior (six behaviors, in particular) as measured in brief thin slices. Instead of asking whether traits predict behavior, our research investigates whether thin slices of behavior predict behavior from a longer behavioral stream. Just as Moskowitz (1986) indicated that some characteristics may be consistent across situations (i.e., situations would not cause variation in the appearance of that characteristic), we extend this logic to thin slices of behavior. To what extent is behavior stable across the longer behavioral stream? Are thin slices of behavior consistent across an interaction? If so, which behaviors? In the present article, stability (consistency) refers to correlations among slices and not the mean level of the behavior between slices. The latter would address how the *amount* of each behavior varied across slices—for example, whether people smile more at the beginning of the interaction than in the middle. Although this too is an interesting question, it is not the subject of the present article, which is only about prediction.

Only one study has directly investigated the relative merits of different-length slices of behavior from a longer behavioral stream for capturing the behavior itself (as opposed, for example, to comparing accuracy of inferring something about the target from different-length slices). Murphy (2005) compared 1-, 2-, and 3-min slices to a longer behavioral stream of 15 min. Trained coders measured target persons' smiling, nodding, gesturing, and self-touching (all by frequency counts) and gazing (by duration) during 15-min dyadic interactions, as well as during three randomly selected 1-min slices from the same interactions. Correlations between the slices and the total (with the respective slices removed from the total) were impressively large, with the average correlation for a single 1-min slice being .68. Furthermore, nothing was gained overall by using two or all three of the slices, meaning that 1 min of behavioral coding was as good as 2 or 3 min. Thus, Murphy's study showed that 1-min slices had good validity, using the full 15 min as the criterion. The present studies were designed to extend Murphy's original study by investigating additional behavior, situations, and slice lengths.

Current Studies

The goals of the present article featuring four studies were to build on previous research and provide guidance for researchers facing methodological decisions with respect to measuring nonverbal behavior. Presently, researchers have little empirical evidence to guide decisions on how much behavior to code and how to sample slices from the longer behavioral stream. We examined six commonly measured nonverbal behaviors: gaze, gestures, nods, self-touch, smiles, and speaking time. Each of these behaviors was measured in at least two of the studies to detect patterns across studies and situations.

Table 1. Reliability and Validity Standards Explored in Studies 1 to 4.

Standard	Question/issue addressed	Measure and statistic
Interslice reliability	How interchangeable are slices in relation to one another?	Intraclass correlation (single measures, two-way mixed, consistency type): ICC(3,1)
Slice–whole validity	How well does a particular slice capture the "whole" interaction for a given behavior?	Corrected item-total correlation: *r*
Cumulative validity	How long a slice is necessary to optimally represent the whole?	Correlation between sums of subsequent slices (from a given behavior) and total for a given behavior: *r*
Behavior validity across slices	Which behaviors are best captured by slices?	Mean corrected item-total correlation across slices within a given behavior: $\bar{r}$
Temporal validity across behaviors	Which slices capture the "whole" interaction best across behaviors?	Mean corrected item-total correlation across behaviors for a given thin-slice time point: $\bar{r}$

Note. ICC = intraclass correlations.

In each study, we measured nonverbal behaviors in thin-slice time intervals. The studies represent a variety of social interaction situations including professional settings and informal, zero-acquaintanceship settings. Following usual practice for behavioral coding, interjudge reliability was always calculated, reflecting agreement among judges who coded the nonverbal slices. The remaining reliability and validity standards investigated across the four studies are summarized below and in Table 1. Broadly speaking, we use the term *reliability* to refer to the stability or consistency of a slice (or slices), whereas *validity* refers to a relationship between the slice and the longer behavioral stream.

Interslice Reliability: How Well, on Average, Do Thin Slices Predict Each Other for a Given Behavior?

Interslice reliability was defined as the intraclass correlation (ICC) across all coded slices for a given behavior. The single-measures ICC represents the average slice reliability for a given behavior, at the level of the individual slice, and is conceptually similar to the mean interslice correlation. Knowing interslice reliability offers the opportunity to apply the Spearman–Brown Prophecy Formula (W. Brown, 1910; Spearman, 1910) to estimate how many slices would need to be added together to reach a desired reliability goal, because reliability of a set of measurements (whether they be items on a test, coders in the laboratory, or slices of behavior within an interaction) is a joint function of the degree of interitem correlation and the number of items (in the present case, slices; Rosenthal, 2005). We illustrate the utility of the Spearman–Brown formula for planning research using the present studies as a guide.

Slice–Whole Validity: How Well do Particular Thin Slices Capture the "Whole" Interaction for a Given Behavior?

Here, "whole" is operationally defined as the total duration of videotaped interaction available for analysis. *Slice–whole validity* is measured in terms of corrected item-total correlations—specifically, the correlation of an individual slice with the sum of the remaining slices for a given behavior. This analysis also revealed whether the temporal location of the slice mattered—for example, does a slice from near the beginning represent the whole for that behavior better or worse than a slice from the middle or end?

Cumulative Validity: How Long a Slice is Necessary to Optimally Represent the Whole?

By summing consecutive slices, we were able to determine the optimal slice length for predicting the total behavior. We calculated cumulative validity by correlating slices of different accumulated lengths with the total amount of a given behavior within the interaction. To calculate cumulative validity, the first slice was correlated with the total behavior, then the first and second slice were summed and that sum was correlated with the total and so forth. This results in longer and longer slices (e.g., 30 s, 1 min, 1.5 min, 2 min) being correlated with the total behavior. Naturally, each cumulative correlation will be positive, because the cumulative slice is included within the behavior total, but this cumulative validity analysis provides information as to what slice lengths (beginning at time zero, the start of the interaction) optimally represent the total behavior in an interaction.

Behavior Validity Across Slices: Are Some Behaviors Better Captured in Thin Slices Than Others?

A researcher might wonder if some behaviors are better or more valid than others to measure as thin slices. *Behavior validity across slices* is defined as the average of the corrected item-total correlation across slices within a given behavior. It allows comparisons between behaviors to

Table 2. Description of Studies 1 to 4 and Behaviors Measured in Each Study.

Study	Interaction type	Interaction length	Slice lengths	Behaviors measured					
				Gaze	Gestures	Nods	Self-touch	Smiles	Speak time
1	Get-acquainted	5 min	30 s	✓	✓	✓	✓	✓	✓
2	Mock job interview	5 min	30 s		✓	✓	✓	✓	
3	Job interview	First 6 min	1 min	✓		✓		✓	✓
4	Medical student with standardized patient	1st, 5th, and 8th min	1 min		✓	✓	✓		

Note. ✓ indicates behavior measured in given study.

determine which behaviors are best represented by slices. It could be the case, for example, that gazing is well captured using thin slices but self-touching is not.

Temporal Validity Across Behaviors: Which Slices Capture the Whole Interaction Best Across Behaviors?

Often, researchers want to be consistent in what slice(s) to use, which means picking a slice (or slices) that work best for all the behaviors they plan to measure, rather than picking different slices for each behavior to be coded. *Temporal validity across behaviors* is therefore the mean corrected item-total correlation across behaviors within a given thin-slice time point. These can then be compared to decide which slice(s) work best for all the behaviors in question.

Description of Studies

Studies 1 and 2 analyzed targets' behavior in a laboratory dyadic interaction, based on 30-s slices, amounting to 5 min total. The interactions in Study 1 were brief "get-to-know-you sessions." Study 2 interactions were mock job interviews with the interviewer and interviewee positions randomly assigned. In Study 1, the slices were coded in sequential order whereas in Study 2, the slices were coded in random order, to rule out the possibility that coding performed sequentially (i.e., judges coded slices in consecutive order) meant that the coder's memory of one slice influenced the coding of subsequent slices. Study 3 involved targets' behavior in a real (not mock) job interview; behavior was extracted in 1-min slices for a total of 6 min. Study 4 investigated three 1-min slices taken from interactions of medical students with an actor playing a patient. Table 2 contains study descriptions and the behaviors measured in each study.

In regard to terminology, throughout the article, we refer to individual slices with the beginning time point (e.g., Slice 0:30 refers to the slice beginning 30 s into the interaction; Slice 1:00 refers to the slice beginning at 1 min 0 s). As mentioned, slices in Studies 1 and 2 were 30 s in length, and those in Studies 3 and 4 were 1 min in length. In all of the studies there were missing observations for individual targets for some slices. To base results on a common set of targets, listwise deletion was used throughout (i.e., a target was not included for a given behavior unless all of the slices were coded for that target).

For ease of interpretation, we will present the methods and procedure of all four studies before presenting the results

Methods and Procedures

Study 1

Target videotaping procedure. College-student targets were videotaped for 5 min in dyads while engaged in a series of interactive tasks. Of those dyads, only the right-seated targets (n = 112; 47 males, 65 females) were coded; only right-seated targets were coded to maintain independence of behavior. Targets were seated at an approximately 45° angle to the camera, facing their interaction partner. Targets were assigned various topics to discuss, including getting acquainted and campus life. Further detail on the original study can be found in Schmid Mast and Hall (2006, Study 2).

Coding procedure and interjudge reliability. Each target's interaction was coded for six nonverbal behaviors; gaze and speaking time were coded as durations whereas gestures, nods, self-touch, and smiles were coded as frequency counts. Three female judges and one male judge were assigned to code one behavior each (gaze, gestures, self-touch, and smiles). Two female judges split the coding for speaking time and two female judges split the coding for nods. For each behavior, interjudge reliability was established between the assigned judge and one other judge by correlating their ratings on ten 30-s slices (10 targets). (Judges were assigned one behavior and each served as the reliability-check judge on one other behavior.) Interjudge reliability was acceptable: gaze r = .97, gestures r = .83, nods r = .79, self-touch r = .67, smiles r = .76, and speaking time r = .88. During coding, judges watched each interaction in 30-s segments, pausing the tape at the end of each segment to record their rating before continuing to the next segment. Only the target was visible during coding; one-half of the screen was covered to conceal the interaction partner. The audio was turned off during coding, with the exception of speaking time.

Study 2

Study 2 again was based on 30-s slices but in this study the judge who measured the thin slices did not code the slices in the order in which the slices occurred on the videotape, but rather in a random order so that temporal effects could not be influenced by memory carryover from one slice to the next. Study 2 examined all of the behaviors examined in Study 1 except for speaking time.

A trained research assistant coded target nonverbal behavior in 30-s slices from a 5-min interaction. Targets were the applicants in a mock job interview where the applicant and interviewer were randomly assigned to their roles (Ruben, Hall, & Schmid Mast, 2014).

Target videotaping procedure. Ninety-eight undergraduates were videotaped in dyads; only the applicants (18 male, 31 female) were used as targets. Targets were seated at approximately a 45° angle to the camera (partially facing their interaction partner and partially facing the camera). Targets were given short descriptions of a newspaper reporting job and their role of applicant; all interviews were 5 min in duration.

Coding procedure and interjudge reliability. The targets' behaviors for the full 5-min interaction were coded, using the same coding methods as in Study 1 (except for the order in which the slices were coded) for gaze, gestures, nods, self-touch, and smiles. To assess reliability, the primary judge coded all 5-min interactions, while a second judge coded 10 targets for each of the five nonverbal behaviors. Interjudge reliability was strong: gaze $r = .97$, gestures $r = .96$, nods $r = .92$, self-touch $r = .99$, and smiles $r = .86$.

After reliability was assessed, each target's full 5-min interview was divided into consecutive 30-s slices, so that each target had a total of 10 slices. The 10 slices were then randomly ordered. One judge coded the 490 target slices, covering the person not of interest and the audio was turned off.

Study 3

The Study 3 interactions were actual job interviews of students applying for a research assistant position. Behavior was analyzed from 1-min slices of the first 6 min of a job applicant being interviewed by a recruiter. (The same recruiter interviewed all applicants and all applicants were asked the same series of questions.) Job applicants were recruited via job announcements placed at various locations at different universities; these job applicants were the targets in Study 3. Before being interviewed for the position, all applicants gave informed consent for being videotaped during the job interview. In Study 3, we examined four behaviors: gaze, smiles, nods, and speaking time. The slices were coded in sequential order.

Target videotaping procedure. Sixty-two mostly undergraduate targets (45 females, 17 males) applying for a research assistant job were videotaped during a job interview (average duration = 11 min; range = 6-19 min). Because the duration of the job interview varied, we based the analysis on the duration of the shortest interview and analyzed the first 6 min of each job interview. The "total" behavior was thus the behavior summed up over the first 6 min.

The applicant (target) was facing the recruiter during the interview and both were seated at a table. Cameras were placed in the middle of the table so that targets' upper bodies were recorded. Before the job interview, the recruiter provided targets with a description of the job for which they were applying (Frauendorfer, Schmid Mast, Nguyen, & Gatica-Perez, 2014b).

Coding procedure and interjudge reliability. Target gaze and smiles were coded by one external judge. Gaze was defined as the duration a target looked at the recruiter and was coded for each of the first 6 minutes of the interaction. Smiles were coded based on a general impression scale ranging from 1 (*not at all smiling*) to 5 (*very much smiling*) for each of the first 6 minutes; the judge made smile ratings after watching each 1-min slice. For both behaviors, the external judge coded all interactions while a second judge coded a sub-sample of five targets. Interjudge reliability was strong: gaze $r = .95$ and smiles $r = .95$.

Target nods and speaking time were automatically coded for each 1-min slice of the first 6 min. Automatic coding means that no human coder was used. Instead, speaking time was extracted with a "microcone," which is a microphone array capable of segmenting speech by speaker (Dev-Audio, 2012). For the assessment of speaking time, the sum of all speaking turn lengths of the applicant was used. Nods were defined as the duration a target nodded while the recruiter was speaking. Nodding was extracted automatically from the videotapes based on a computer algorithm developed and validated for this video material (see also Frauendorfer, Schmid Mast, Nguyen, & Gatica-Perez, 2014a; Nguyen, Odobez, & Gatica-Perez, 2012).

Study 4

Medical students' behavior was measured during a clinical interaction with an actor-patient in three 1-min slices from the first, fifth, and ninth minutes of the interaction (corresponding to Slice 0:00, Slice 4:00, and Slice 8:00 in the interaction; Hall, Roter, Blanch, & Frankel, 2009a, 2009b). These 3 min were the only minutes coded, so the combination of these 3 min was called "total" behavior for the present analyses.

Target videotaping procedure. Third-year medical students were videotaped in an interaction with a standardized (actor) patient as part of a required evaluation of clinical skills

(average interview duration was approximately 15 min). The overall sample size was 145 (79 male, 66 female), though deletions due to technical issues and missing data brought the *N* down depending on the analysis. The camera position did not allow measurement of some nonverbal cues (e.g., gaze and smiles).

Coding procedure and interjudge reliability. Slices 0:00, 4:00, and 8:00 (1-min duration each) were coded. Gestures, nods, and self-touch were coded in the same way as Studies 1 and 2. Judges were trained to achieve interjudge reliability of r = .70 against an independent coder.

Results

For ease of interpretation, we present results according to the type of reliability or validity assessed, and present each behavior as measured across studies. Study 4 data are not plotted in Figures 2, 3, and 5 because Study 4 slices were not equivalent to the timelines of Studies 1 to 3; however, Study 4 data are presented and discussed in conjunction with Studies 1 to 3 within the text.

Interslice Reliability

Interslice reliability was measured to reflect how well slices predict one another on average, for a given behavior. ICCs [3,1] (Shrout & Fleiss, 1979) were calculated for all slices within a given behavior, across each study. The results are presented in Figure 1. Gaze had impressive interslice reliability in all of the studies in which it was measured, and nods were also consistent but weaker in magnitude. Other behaviors were not so consistent: smiles, gestures, and self-touch all had very good interslice reliability in one study each, but in other studies these behaviors did not show as good interslice reliability. Speaking time never showed good interslice reliability.

Application of the Spearman–Brown formula enables a researcher to project how many slices would be needed to achieve a given reliability for those slices taken together (e.g., Cronbach's α). This exercise would be useful for a researcher who wishes to avoid coding too few slices or needs to defend whatever choice was made against criticism. The formula is reliability $= \left(n\bar{r}\right)/1+\left(n-1\right)\bar{r}$, where n refers to the number of slices (or items or coders, in other applications) and $\bar{r}$ refers to the average interslice (or interitem or intercoder) correlation. To illustrate based on the current studies, for gaze with an approximate interslice reliability of .60, it would take only three slices to achieve an effective reliability of .80 or more (e.g., Cronbach's α; Rosenthal, 2005). In the case of nods, with an approximate interslice reliability of .40, it would take six slices to achieve the same reliability, and in the case of speaking time with an approximate interslice reliability of .25, it would take 12 slices. On the other hand, if one would accept an overall reliability of .60, then five slices would suffice for speaking time (in each example, we refer to slices of the same durations as those used in the present studies).

Slice–Whole Validity

How well do particular thin slices capture the "whole" interaction for a given behavior? Slice–whole validity was measured by calculating the item-total correlation for each slice for a given behavior. Figure 2 presents the results according to behavior across studies, with the exception of Study 4. (Supplementary Tables S1 through S4 contain the exact item-total correlation values as well as the 95% confidence intervals [CIs] for each correlation in all studies.)

Gaze showed the highest consistency across the three studies that measured gaze. Most gaze item-total correlations were above .65. Nods were measured in all four studies and showed some slice–whole validity consistency across studies and slices, particularly at later slices, though the magnitude of the validity was smaller than for gaze. For nods, slice–whole validity was generally above .50. Study 4 nod (not plotted in Figure 2) slice–whole validity values were Slice 0:00 r = .56, Slice 4:00 r = .59, and Slice 8:00 r = .54, which replicated the findings in the other three studies. Across the three studies that measured smiles, there was fair consistency for slice–whole validity, with notably exceptions occurring at the latest slices. Slice–whole validity for smiles was generally above .50, and the highest values were obtained in Study 3.

Self-touch in Studies 1 and 2 typically showed slice–whole validity values above .60, but Study 2 showed a noticeable dip in self-touch slice–whole validity at later slices. However, the results from Study 4 (not plotted in Figure 2) demonstrated much lower self-touch slice–whole validity with Slice 0:00 r = .08, Slice 4:00 r = .33, and Slice 8:00 r = .42. Unlike Studies 1 and 3, slice–whole validity values increased incrementally with each successive slice in Study 4. The lowest self-touch slice–whole validity values for all three studies occurred at Slice 0:00.

Slice–whole validity for gestures and speaking time was much less consistent across the studies that included these behaviors. For gestures, there was relatively little overlap between the values across slices in Studies 1 and 2, as shown in Figure 2. Gestures slice–whole validity in Study 4 (not plotted in Figure 2) had comparatively lower values than Studies 1 and 2 with Slice 0:00 r = .27, Slice 4:00 r = .33, and Slice 8:00 r = .34. Speaking time had the lowest slice–whole validity values, which tended to be below .60 and in many cases, below .50. And similar to gestures, the slice–whole validity pattern did not appear consistent across studies with this behavior.

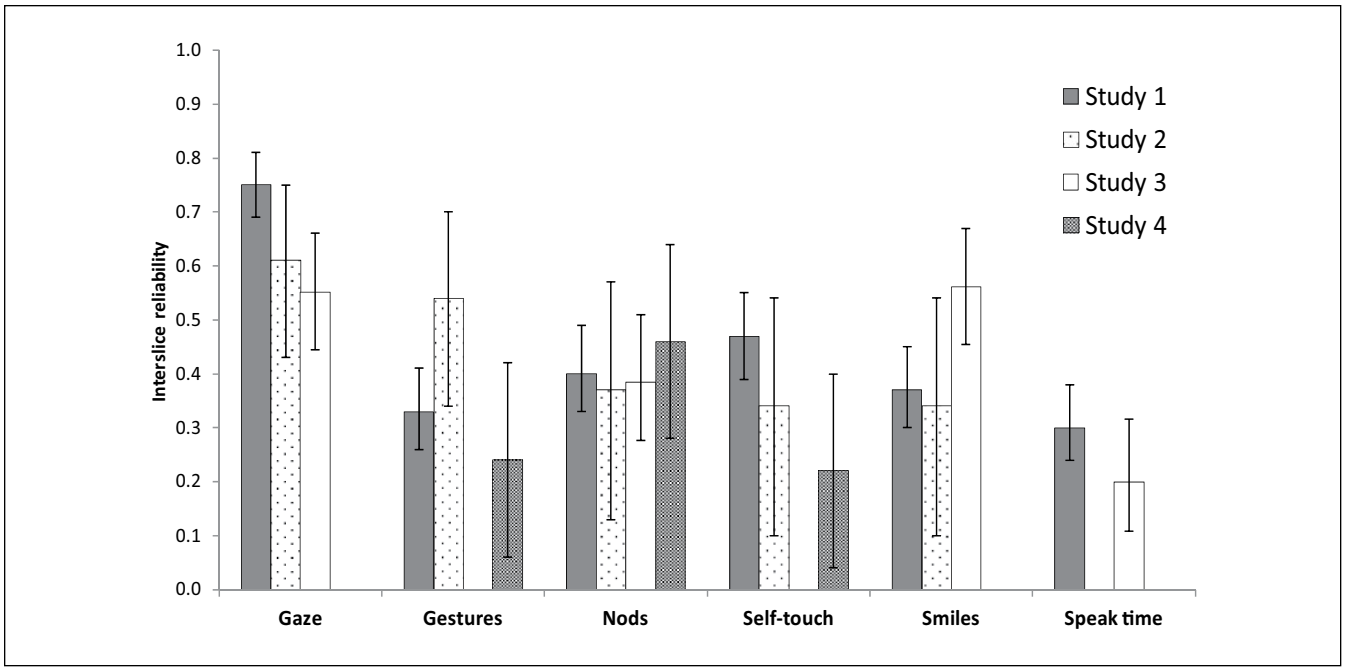

Figure 1. Interslice reliability by behavior.

Note. Each bar represents the intraclass correlation (ICC; single measures, two-way mixed, consistency type) of all slices measured for a given behavior within each study. For Studies 1 and 2, behavior was measured in 30-s slices across 5 min, resulting in 10 slices. For Study 3, behavior was measured in 1-min slices across 5 min, resulting in 5 slices. For Study 4, three 1-min slices were coded across approximately 15 min (corresponding to the 1st, 5th, and 9th min of the interaction). Study 1 *n*s ranged from 101 to 111; Study 2 $n = 40$; Study 3 $n = 62$; Study 4 *n*s ranged from 92 to 106. Error bars represent 95% confidence intervals.

Cumulative Validity

By summing consecutive slices, we could ask what is the optimal slice length that predicts the total behavior? We calculated cumulative validity by correlating increasingly long slice durations with the total amount of a given behavior within the interaction. This was done by correlating the first slice with the behavior total, then summing the first and second slice and correlating the sum with the behavior total, and continuing correlating summed subsequent slice sums with the behavior total. Figure 3 displays the results for behaviors in Studies 1 to 3. The general pattern across all six behaviors was fairly consistent. Not surprisingly, the correlations got incrementally larger (and closer to the total behavior) with each additional slice summed. This pattern was repeated in Study 4 (which is not plotted in Figure 3) with the correlations between Slice 0:00 and behavior totals as follows: gesture $r = .56$, nods $r = .73$, and self-touch $r = .48$. And the correlations between the sum of Slice 0:00 + Slice 4:00 with behavior totals increased to gesture $r = .86$, nods $r = .91$, and self-touch $r = .91$. Note that for Study 4, behavior totals were based on the sum of three 1-min slices, thus the sum of Slice 0:00 + Slice 4:00 represented two thirds of the behavior total. Across all studies, validity values achieved $r > .70$ within the first 1.5 to 2 min of an interaction, with the exception of speaking time.

Behavior Validity Across Slices

Behavior validity across slices measured which behaviors are best represented by a thin slice, on average, regardless of which slice of the interaction is selected. For each study, behavior validity across slices was calculated by averaging the corrected item-total correlations across slices within a given behavior. The item-total correlations were transformed to Fisher-*z* before averaging and then converted back to the *r* metric for presentation. As an example, slice–whole validity for gaze in Study 1 was calculated by averaging the item-total correlations plotted in the upper-left panel of Figure 2. (Supplementary Tables S1 through S4 contain the exact item-total correlation values, as well the 95% CIs for each correlation, on which the behavior validity by slice calculations were based.)

The results for all six coded behaviors across studies are shown in Figure 4. Gaze showed the strongest and most consistent behavior validity with values above .69 in the three studies that measured gaze. Nods also showed remarkably consistent behavior validity, albeit lower in magnitude compared to gaze. While slightly more variable across three studies, smiles also showed relatively strong behavior validity (with the lowest value $\bar{r} = .53$ in Study 2). These results indicate that these behaviors, among those studied here, are the ones that best represent the total. Gestures, self-touch, and

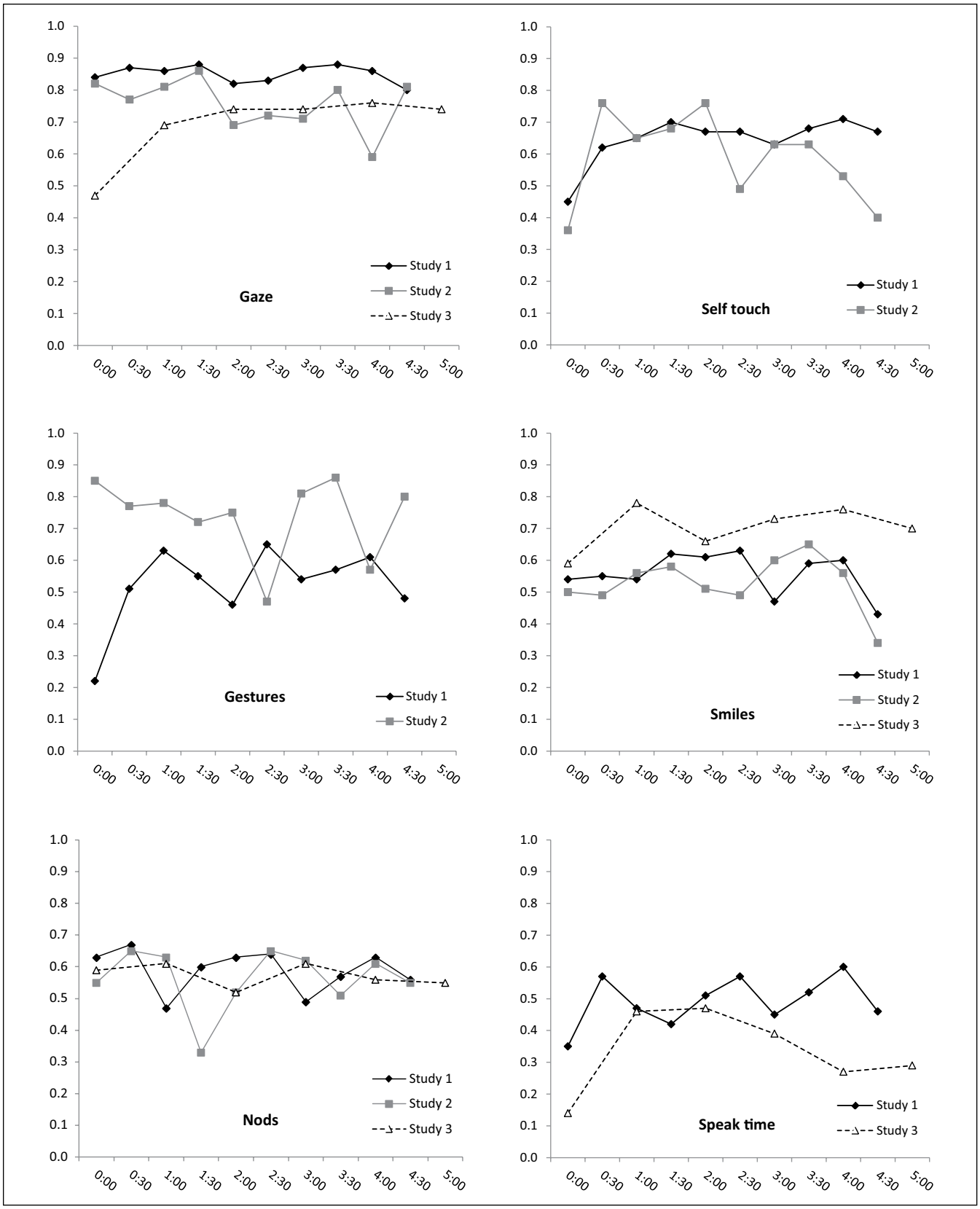

Figure 2. Slice–whole validity, measured as item-total correlations (*y* axes) for each slice (*x* axes) by behavior.
Note. Each data point represents the corrected item-total correlation (*r*) between the behavior in the given slice as correlated with the total behavior across the entire interaction. For Studies 1 and 2, behavior was measured in 30-s slices; for Study 3, behavior was measured in 1-min slices. Study 1 *n*s ranged between 101 and 111, Study 2 $n = 40$, Study 3 $n = 62$.

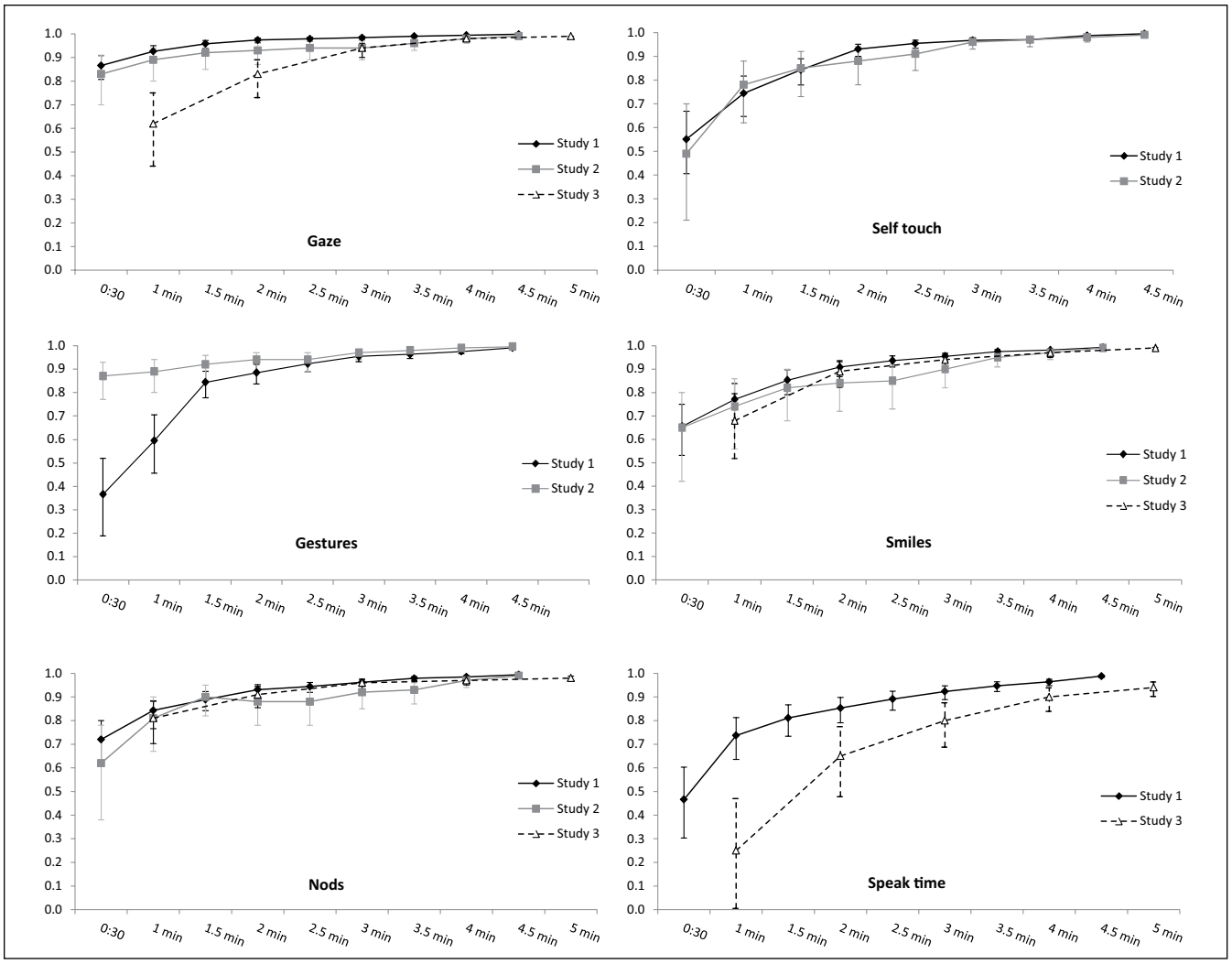

Figure 3. Cumulative validity for each behavior across slices.

Note. Cumulative validity was calculated by correlating the total behavior of the interaction with cumulative slice lengths. Thus, the first slice was correlated with the behavior total, then the first and second slice were summed and correlated with the behavior total, and so forth. The *y* axes represent the cumulative slice–behavior total correlations (*r*) and the *x* axes represent the cumulative slice length (e.g., 2.5 min represents the first 2.5 min of the interaction). For Studies 1 and 2, behavior was measured in 30-s slices; for Study 3, behavior was measured in 1-min slices. Study 1 *n*s ranged between 101 and 111, Study 2 $n = 40$, Study 3 $n = 62$. Errors bars represent 95% confidence intervals.

speaking time showed inconsistent behavior validity values across studies. Study 4 contained the lowest behavior validity values for both gestures and self-touch.

Temporal Validity Across Behaviors

In assessing temporal validity across behaviors, we asked which slices best captured the entire interaction, regardless of behavior. For this analysis, we averaged the corrected item-total correlations across behaviors for each slice within each study. As in the previous analyses, the item-total correlations were transformed to Fisher-*z* before averaging and then converted back to the *r* metric for presentation. As an example, temporal validity across behaviors for Slice 0:00 of Study 1 was calculated by averaging the Slice 0:00 item-total correlations for all six behaviors of Study 1. (Supplementary Tables S1 through S4 contain the exact item-total correlation values, as well the 95% CIs for each correlation, on which the temporal validity across behavior calculations were based.) The results are plotted in Figure 5.

The pattern is fairly striking; averaging across behaviors, Figure 5 suggests that no one slice is superior to another in terms of temporal validity. Temporal validity was substantially positive and generally similar in magnitude across slices, indicating substantial consistency from slice to slice. Though lower in magnitude, Study 4 (not shown in Figure 5) demonstrated a similar pattern with temporal validity across behaviors for Slice 0:00 $\bar{r} = .32$, Slice 4:00 $\bar{r} = .43$, and Slice 8:00 $\bar{r} = .44$. However, Slice 0:00 had the lowest temporal validity across behaviors in three of the four studies. This pattern suggests that the first slice of the interaction may not best represent the longer interaction.

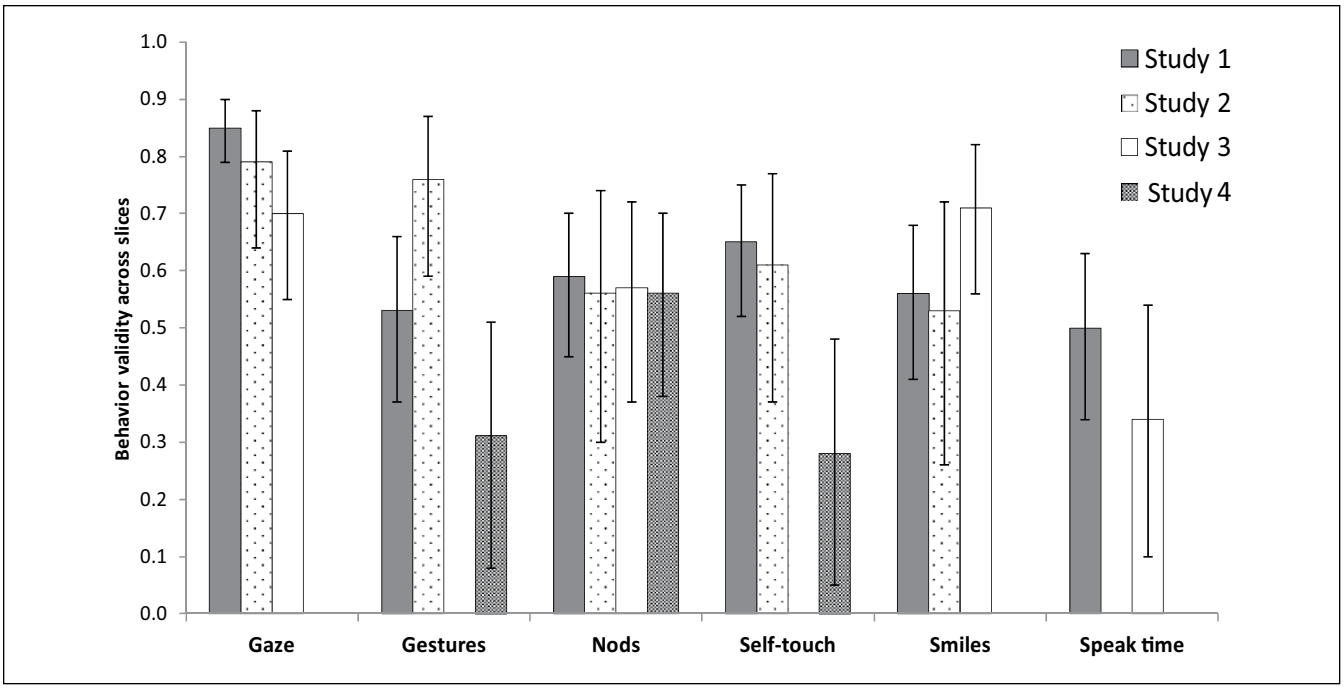

Figure 4. Behavior validity across slices.
Note. Each bar represents the mean corrected item-total correlation ($\bar{r}$) across slices for the given behavior for each study. Study 1 *n*s ranged from 101 to 111; Study 2 $n = 40$; Study 3 $n = 62$; Study 4 *n*s ranged from 92 to 106. Error bars represent 95% confidence intervals.

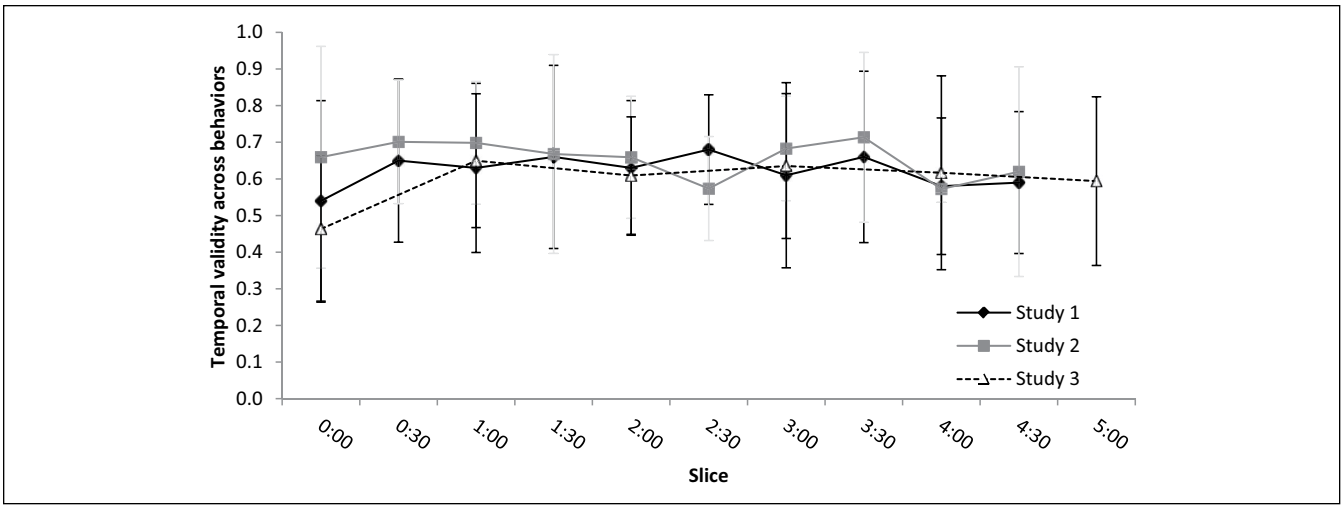

Figure 5. Temporal validity across slices.
Note. Each data point represents the mean corrected item-total correlation ($\bar{r}$) across all behaviors for a given slice in each study (Studies 1-3). Study 1 *n*s ranged from 101 to 111; Study 2 $n = 40$; Study 3 $n = 62$. Error bars represent 95% confidence intervals.

Discussion

This article investigated the reliability and validity of thin slices from various perspectives. As researchers who study nonverbal behavior in social interactions, we were interested in how well thin slices of nonverbal behavior capture the entirety of behavior across the whole interaction. Given the arduous nature of nonverbal coding, we were also interested in information and techniques that might reduce the labor involved in behavior coding. The four studies investigated thin slices of six commonly measured nonverbal behaviors, both 30-s and 1-min slices, from interactions of various types and lengths. Results across four studies provide substantial empirical evidence for methodological decisions about

coding nonverbal behavior and have potential implications regarding behavioral consistency in social interactions. Below we summarize the results and compare the findings with previous literature.

Interslice Reliability

Interslice reliability tested how interchangeable the various slices are with one another. That is, how well do slices predict each other? Figure 1 displayed these results. Almost all behaviors demonstrated significant interslice reliabilities. Gaze and nods demonstrated the strongest and most consistent interslice reliability, though gestures, self-touch, and smiles had good interslice reliability in one study each. These results indicate that individual thin slices of these behaviors predicted other individual slices of the same behavior reasonably well on average.

On the other hand, the weakest interslice reliability occurred for speaking time, particularly in Study 3. Perhaps speech-related behaviors such as speaking time may be strongly dependent on the interaction partner's behavior rather than representing the target's personal style of communicating nonverbally. Just as K. W. Brown and Moskowitz (1998) found certain traits were displayed more consistently than others (e.g., dominance compared to agreeableness), we found that certain behaviors were displayed more consistently than others; notably, gaze stood out in terms of consistency.

We also showed that, using correlations among slices, one can estimate the overall reliability of any number of slices using the Spearman–Brown Prophecy Formula. Researchers could conduct this analysis in a post-hoc manner for the number of slices they used, or they could do it for planning purposes based on pilot data or past studies. They could discover either that fewer slices were or are needed than they thought, or that more slices are needed than they thought.

Slice–Whole Validity Patterns

Slice–whole validity, measured as item-total correlations, tested how well a particular slice captured the whole interaction for a given behavior (as shown in Figure 2). Gaze showed the strongest patterns, with relatively consistent slice–whole validity across slices and studies. The results suggest that 30-s or 1-min slices extracted from anywhere in the interaction may adequately represent gazing behavior, with strongest convergence across three studies in the middle slices (Slices 2:00 to 3:00).

Nods and smiles also showed relatively consistent patterns of slice–whole validity, but at a smaller magnitude (in comparison to gaze). Across three studies, higher consistency (where values converged across studies) occurred at later slices for nods and earlier slices for smiles. The remaining behaviors of gestures, self-touch, and speaking time were much more variable in slice–whole validity across studies. Gestures, in Studies 1 and 2, had a wide range of slice–whole validities, and had weak slice–whole validity in Study 4 (below .35), suggesting that gesture slices of 30-s or 1-min durations may not adequately represent the entirety of the interaction. Because gesturing is typically associated with speech, this finding may reflect dyadic interaction more than an individual behavior style. Similar results were found for self-touch, and the range and values of slice–whole validity across studies suggest that slices of self-touch should be carefully considered as to whether these adequately represent the entire interaction. Self-touch is often indicative of anxiety and perhaps anxiety levels fluctuate throughout a social interaction. Speaking time had some of the lowest slice–whole validity, again, suggesting that 30-s or 1-min slices may not adequately represent speaking time across the whole interaction. While the smaller and more variable slice–whole validity values for gestures, self-touch, and speaking time may imply that slices may not adequately represent these behaviors, we are disinclined to suggest that slices of these behaviors should not be used, as some values were quite substantial (e.g., slice–whole validity for gestures in Study 2 was $\geq .70$ for 30-s slices from time 0:00 to 2:00). Some of the weakest slice–whole validity values were found in Study 4, which may be partially explained by the fact that only three slices were measured and therefore the corrected total consisted of only two slices, a situation that would undoubtedly reduce reliability of the corrected total relative to the other studies where more slices were aggregated into the corrected total, which in turn would impact the extent to which individual slices could correlate with it.

In general, the slice–whole validity results are remarkably consistent with Murphy's (2005) findings for 1-min slices of five behaviors: gaze, gestures, nods, self-touch, and smiles. As in the present studies, Murphy found that slices of gaze and nods showed strong and consistent results, and self-touch was the weakest. (Speaking time was not measured in the Murphy study.)

Cumulative Validity Patterns

To investigate the optimal slice length that predicts the total behavior displayed in an interaction, we investigated the correlations between behavior in cumulative slices and the total behavior (Figure 3). Across the four studies, most behaviors achieved strong cumulative validity within the first 1.30-2.00 slice length. Just as more raters substantially increased the generalizability of ratings (Moskowitz & Schwarz, 1982), we found that increases in the number of slices increased the cumulative validity, and the asymptotes were reached rather quickly (within 1-2 min). The data clearly show that it is not necessary to code the entire duration of behavior to have equivalent cumulative validity; the first 1.30-2.00 min of interactions appear to adequately represent the behaviors of gaze, gestures, nods, self-touch, and smiles. Given that the first minute had the lowest cumulative validity

values, perhaps accumulating the first 1:30-2:00 min is not optimal. Researchers might consider beginning coding later in the social interaction, especially because middle slices, in several cases, had the highest slice–whole validity values.

Speaking time was the exception; it took between 3:30 and 4:00 min to achieve sufficient cumulative validity. The two studies that measured speaking time were dissimilar (unstructured get-to-know-you interactions of Study 1 compared to genuine job interviews in Study 3) and perhaps speaking time is more situationally dependent. At the very least, the speaking time cumulative validity results suggest that longer slices of an interaction should be used if the slice is to represent speaking time across the entire interaction.

Patterns of Behavior Validity Across Slices and Temporal Validity Across Behaviors

Behavior validity across slices measured which behaviors are best represented by a thin slice, on average, regardless of which slice of the interaction is selected (Figure 4). Similar to the slice–whole validity patterns, gaze showed the strongest and most consistent behavior validity, with nods and smiles also showing satisfactory behavior validity across slices. Of the six nonverbal behaviors measured in the present studies, these three behaviors were best captured by slices. Maybe gaze, nods, and smiles are more personal habits/styles than some other behaviors, unlike speaking time, for instance, which may depend a lot on the interaction partner.

Gestures, self-touch, and speaking time showed more variability across studies, indicating that these behaviors may be more situationally dependent. As mentioned, speaking time may be more dependent on the interaction partner, and because gestures are closely linked to speaking turns, these behaviors are more variable across an interaction. Likewise, self-touch, which typically reflects anxiety, may also be more variable across an interaction and situationally dependent. Study 4 contained the lowest behavior validity values for gestures and self-touch, with values much lower than found in the other studies that measured these behaviors. Perhaps behavior validity was particularly low in Study 4 due to the situational context of a medical student interacting with an actor-patient. In such a structured interaction, perhaps nonverbal behavior is displayed less consistently than in the other, more naturalistic or unstructured settings. Study 4 also likely demonstrated low behavior validity because it was based on only three 1-min slices for all measured behaviors, whereas the other studies had more slices in the calculations.

One implication suggested by the temporal validity analyses is that a brief slice from the very beginning of an interaction may not be the best indicator of behavioral patterns from the entire interaction. Perhaps this pattern reflects an early hypothesis by Goffman (1963), who suggested that the beginning of an interaction is somewhat scripted and thus may not indicate individual behavioral variation. At the very least, the present results signal that the first slice of the interaction may not best represent the longer interaction.

Implications

There are practical and theoretical implications of these results. At a practical level, the findings suggest that, regardless of behavior, 30-s or 1-min slices from after the first minute of an interaction may adequately represent certain nonverbal behaviors, particularly for gaze, nods, and smiles. Using the first 1.5 to 2 min of an interaction to represent behavior from the whole interaction is sufficient for several behaviors, but longer slices to represent speaking time are surely needed. Combined, the results suggest that slices after the first minute, and before the last minute, are a researcher's best bet.

Although the present results are very encouraging with regard to the reliability and validity of 30-s or 1-min behavioral slices, it does not necessarily follow that researchers should use slices that are short. While in many cases, the results for 30-s or 1-min slices were as good as those for 1, 2, and even 3 min in Murphy's (2005) study, it should be remembered that generally overall reliability and validity did go up as more slices were aggregated. When making decisions about nonverbal measurement, researchers will undoubtedly consider a number of factors, including the minimum reliability and validity acceptable to them, the number of behaviors to be measured, the number of targets, and the resources available. To the extent that direct behavioral observation may be one of the strongest methods for measuring behavior (Furr, 2009), the results presented here suggest that thin-slice behavioral measurement may be a sufficient coding strategy.

At a theoretical level, these data may relate to larger questions about behavioral consistency and personality. In the personality literature, previous doubts were raised about the temporal reliability of a construct, with researchers emphasizing aggregation: "It is important to sample behavior over stimuli and situations in order to generalize over stimuli and situations" (Epstein, 1983a, p. 364). Moskowitz (1986) indicated that, to demonstrate the convergence between a behavioral measure and other measures (i.e., personality traits), a sufficient number of occasions should be sampled to generalize across situations. However, the present results imply fair amounts of high cross-situational consistency for several behaviors measured in thin slices, particularly because behaviors were measured in several studies involving various types of social interaction. And considering the relatively high slice–whole validity and cumulative validity values, our results suggest that thin slices can reliably measure behavior and adequately represent a longer behavioral stream. While the present work did not address whether the six measured behaviors indicate global personality characteristics—and much of the behavioral consistency–personality debate has been about that very question—we believe the four studies here provide an empirical starting point from which personality and nonverbal cues could be further investigated. We echo previous appeals that researchers consider how well the measured behavior fits the construct of interest (Blackman & Funder,

1998; Epstein, 1983b; Moskowitz, 1986)—what is the purpose of measuring the behavior? While one behavior in isolation may reveal some information about a person's internal thoughts, feelings, or traits, most behaviors are processed in conjunction with other behaviors and the context (Murphy, 2012; Patterson, 1995). Thus, researchers should consider the whole picture (i.e., context, behaviors, personality constructs) when making decisions about behavioral coding.

Limitations and Conclusions

Of course, our findings are qualified by several considerations. To begin with, we only measured six nonverbal behaviors, which were highly observable behaviors. For example, in Study 1, the average behavior totals were 128 s of gaze, 19 gestures, 14 nods, 33 self-touches, 15 smiles, and 102 s of speaking time in the 5-min interactions. Thus, these behaviors were displayed with some frequency throughout the interaction. As indicated by Moskowitz and Schwarz (1982), base rates of behavior affect reliable measurement of a given behavior, with less frequent behaviors requiring longer observations. Thus, generalization of the findings to less frequent behaviors (e.g., forward leaning, crossed/closed arms) warrants caution, as these behaviors may be less reliably assessed.

Related to frequency, the abstraction level of the coded behavior or category should be considered. All behaviors measured in the present studies have concrete levels of abstraction in that the behaviors could be quantified as time or frequency. However, "fuzzy" constructs with higher levels of abstraction (e.g., "pleasant speech style") may not demonstrate such high levels of validity or reliability (Epstein, 1983a). Yet, whether the construct is concrete or "fuzzy," the question of how much of an interaction to code would remain the same and we believe the present results are at least a starting point in that direction. And all extrapolations about thin slices representing longer behavioral streams must be tempered by the fact that the longest of any of our interactions from which slices were selected was about 15 min in length. It is unknown whether these results would extend to longer interactions or different types of interactions. Of course, each of these issues is an empirical question and future research could investigate the reliability or validity of thin slices with regard to behavior frequency, abstract constructs, and interaction length.

In sum, the present results should give researchers considerable confidence in the thin-slice approach for behavioral coding. There is renewed interest and advocacy in measuring behavior, particularly through direct observation, within personality psychology and the field of psychology as a whole (Back & Egloff, 2009; Baumeister, Vohs, & Funder, 2007; Furr, 2009; Patterson, 2014). We hope our article represents a step in this direction and provides guidance for investigators who are interested in using thin slices to represent various behaviors across longer behavioral streams.

Acknowledgments

We thank the many judges whose labor produced the thin-slice behavior data analyzed in this article.

Authors' Note

Mollie A. Ruben is now at the Center for Healthcare Organization and Implementation Research, U.S. Department of Veterans Affairs. Danielle Blanch-Hartigan is now at the Department of Natural and Applied Sciences, Bentley University.

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: The Fetzer Institute and the Swiss National Science Foundation (CRSII2_147611).

Supplemental Material

The online supplemental material is available at http://pspb.sagepub.com/supplemental.

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. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 32, pp. 201-271). San Diego, CA: Academic Press.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*, 256-274. doi:10.1037/0033-2909.111.2.256

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. doi:10.1037/0022-3514.64.3.431

Back, M. D., & Egloff, B. (2009). Yes we can! A plea for direct behavioral observation in personality research. *European Journal of Personality*, *23*, 403-405. doi:10.1002/per.725

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*, 396-403. doi:10.1111/j.1745-6916.2007.00051.x

Blackman, M. C., & Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. *Journal of Experimental Social Psychology*, *34*, 164-181. doi:10.1006/jesp.1997.1347

Borkenau, P., Brecke, S., Möttig, C., & Paelecke, M. (2009). Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, *43*, 702-705. doi:10.1016/j.jrp.2009.03.007

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*, *86*, 599-614. doi:10.1037/0022-3514.56.3.398

Brown, K. W., & Moskowitz, D. S. (1998). Dynamic stability of behavior: The rhythms of our interpersonal lives. *Journal of Personality*, *66*, 105-134. doi:10.1111/1467-6494.00005

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296-322.

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, *41*, 1054-1072. doi:10.1016/j.jrp.2007.01.004

Dev-Audio [Microcone apparatus]. (2012). Gold Coast, Australia. Retrieved from http://www.dev-audio.com/products/microcone

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, *37*, 1097-1126. doi:10.1037/0022-3514.37.7.1097

Epstein, S. (1983a). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, *51*, 360-392. doi:10.1111/j.1467-6494.1983.tb00338.x

Epstein, S. (1983b). The stability of confusion: A reply to Mischel and Peake. *Psychology Review*, *90*, 179-184. doi:10.1037/0033-295X.90.2.179

Fairbairn, C. E., Sayette, M. A., Levine, J. M., Cohn, J. F., & Creswell, K. G. (2013). The effects of alcohol on the emotional displays of Whites in interracial groups. *Emotion*, *13*, 468-477. doi:10.1037/a0030934

Frauendorfer, D., Schmid Mast, M., Nguyen, L. S., & Gatica-Perez, D. (2014a). Nonverbal social sensing in action: Unobtrusive recording and extracting of nonverbal behavior in social interactions illustrated with a research example. *Journal of Nonverbal Behavior*, *38*, 231-245. doi:10.1007/s10919-014-0173-5

Frauendorfer, D., Schmid Mast, M., Nguyen, L. S., & Gatica-Perez, D. (2014b). *A step towards automatic applicant selection: Predicting job performance based on applicant nonverbal interview behavior*. Manuscript in preparation.

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations and behaviors. *Journal of Research in Personality*, *40*, 21-34. doi:10.1016/j.jrp.2005.08.003

Funder, D. C. (2009). Persons, behaviors and situations: An agenda for personality psychology in the postwar era. *Journal of Research in Personality*, *43*, 120-126. doi:10.1016/j.jrp.2008.12.041

Furr, R. M. (2009). The study of behaviour in personality psychology: Meaning, importance, and measurement. *European Journal of Personality*, *23*, 437-453. doi:10.1002/per.726

Goffman, E. (1963). *Stigma: Notes on the management of spoiled identity*. New York, NY: Simon & Schuster.

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of Personality and Social Psychology*, *74*, 1337-1349. doi:10.1037/0022-3514.74.5.1337

Grahe, J. E., & Bernieri, F. J. (1999). The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior*, *23*, 253-269. doi:10.1023/A:1021698725361

Hall, J. A., Andrzejewski, S. A., Murphy, N. A., Schmid Mast, M., & Feinstein, B. A. (2008). Accuracy of judging others' traits and states: Comparing mean levels across tests. *Journal of Research in Personality*, *42*, 1476-1489. doi:10.1016/j.jrp.2008.06.013

Hall, J. A., Carter, J. D., & Horgan, T. G. (2001). Status roles and recall of nonverbal cues. *Journal of Nonverbal Behavior*, *25*, 79-100. doi:10.1023/A:1010797627793

Hall, J. A., Roter, D. L., Blanch, D. C., & Frankel, R. M. (2009a). Nonverbal sensitivity in medical students: Implications for clinical interactions. *Journal of General Internal Medicine*, *24*, 1217-1222. doi:10.1007/s11606-009-1107-5

Hall, J. A., Roter, D. L., Blanch, D. C., & Frankel, R. M. (2009b). Observer-rated rapport in interactions between medical students and standardized patients. *Patient Education and Counseling*, *76*, 323-327. doi:10.1016/j.pec.2009.05.009

Hall, J. A., Roter, D. L., & Rand, C. S. (1981). Communication of affect between patient and physician. *Journal of Health and Social Behavior*, *22*, 18-30.

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist*, *43*, 23-34. doi:10.1037/0003-066X.43.1.23

Leary, M. R., & Hoyle, R. H. (2009). Situations, dispositions, and the study of social behavior. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 3-11). New York, NY: Guilford.

Matsumoto, D., LeRoux, J., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., . . .Goh, A. (2000). A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian Brief Affect Recognition Test (JACBART). *Journal of Nonverbal Behavior*, *24*, 179-209. doi:10.1023/A:1006668120583

Milmoe, S., Rosenthal, R., Blane, H. T., Chafetz, M. E., & Wolf, I. (1967). The doctor's voice: Postdictor of successful referral of alcoholic patients. *Journal of Abnormal Psychology*, *72*, 78-84.

Mischel, W. (1968). *Personality and assessment*. New York, NY: Wiley.

Moskowitz, D. S. (1986). Comparison of self-reports, reports by knowledgeable informants, and behavioral observation data. *Journal of Personality*, *54*, 294-317. doi:10.1111/j.1467-6494.1986.tb00396.x

Moskowitz, D. S., & Schwarz, J. C. (1982). Validity comparison of behavior counts and ratings by knowledgeable informants. *Journal of Personality and Social Psychology*, *42*, 518-528. doi:10.1037/0022-3514.42.3.518

Murphy, N. A. (2005). Using thin slices for behavioral coding. *Journal of Nonverbal Behavior*, *29*, 235-246. doi:10.1007/s10919-005-7722-x

Murphy, N. A. (2007). Appearing smart: The impression management of intelligence, person perception accuracy, and behavior in social interaction. *Personality and Social Psychology Bulletin*, *33*, 325-339. doi:10.1177/0146167206294871

Murphy, N. A. (2012). Nonverbal perception. In S. T. Fiske & C. N. Macrae (Eds.), *Handbook of social cognition* (pp. 196-215). London, England: SAGE.

Newton, T. L., Haviland, J. M., & Contrada, R. J. (1996). The face of repressive coping: Social context and the display of hostile expressions and social smiles. *Journal of Nonverbal Behavior*, *20*, 3-22. doi:10.1007/BF02248712

Nguyen, L. S., Odobez, J. M., & Gatica-Perez, D. (2012, October). *Using self-context for multimodal detection of head nods in face-to-face interactions*. Paper presented at the International Conference on Multimodal Interactions, Santa Monica, CA.

Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of NonVerbal Accuracy Scale. *Journal of Nonverbal Behavior*, *18*, 9-34. doi:10.1007/BF02169077

Patterson, M. L. (1995). A parallel process model of nonverbal communication. *Journal of Nonverbal Behavior*, *19*, 3-29. doi:10.1007/BF02173410

Patterson, M. L. (2014). Reflections on historical trends and prospects in contemporary nonverbal research. *Journal of Nonverbal Behavior*, *38*, 171-180. doi:10.1007/s10919-013-0171-z

Rosenthal, R. (2005). Conducting judgment studies: Some methodological issues. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 199-234). New York, NY: Oxford University Press.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore, MD: The Johns Hopkins University Press.

Roter, D. L., Hall, J. A., Blanch-Hartigan, D., Larson, S., & Frankel, R. M. (2011). Slicing it thin: New methods for brief sampling of medical dialogue. *Patient Education and Counseling*, *82*, 410-419. doi:10.1016/j.pec.2010.11.019

Ruben, M. A., Hall, J. A., & Schmid Mast, M. (2014). Smiling in a job interview: When less is more. *The Journal of Social Psychology*. Advance online publication. doi:10.1080/00224545.2014.972312

Rule, N. O., & Ambady, N. (2008). Brief exposures: Male sexual orientation is accurately perceived at 50 ms. *Journal of Experimental Social Psychology*, *44*, 1100-1105. doi:10.1016/j.jesp.2007.12.001

Schmid Mast, M., & Hall, J. A. (2006). Women's advantage at remembering others' appearance: A systematic look at the why and when of a gender difference. *Personality and Social Psychology Bulletin*, *32*, 353-364. doi:10.1177/0146167205282150

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428. doi:10.1037/0033-2909.86.2.420

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271-295.

Tellegen, A. (1991). Personality traits: Issues of definition, evidence, and assessment. In W. M. Grove & D. Cicchetti (Eds.), *Thinking clearly about psychology: Vol. 2. Personality and psychopathology* (pp. 10-35). Minneapolis: University of Minnesota.

Tucker, J. S., & Anders, S. L. (1998). Adult attachment style and nonverbal closeness in dating couples. *Journal of Nonverbal Behavior*, *22*, 109-124. doi:10.1023/A:1022980231204