A novel rater agreement methodology for language transcriptions: evidence from a nonhuman speaker

Allison B. Kaufman, Erin N. Colbert-White & Robert Rosenthal

Quality & Quantity
International Journal of Methodology

ISSN 0033-5177

Qual Quant
DOI 10.1007/s11135-013-9894-5

Springer

Qual Quant
DOI 10.1007/s11135-013-9894-5

A novel rater agreement methodology for language transcriptions: evidence from a nonhuman speaker

Allison B. Kaufman · Erin N. Colbert-White · Robert Rosenthal

Abstract The ability to measure agreement between two independent observers is vital to any observational study. We use a unique situation, the calculation of inter-rater reliability for transcriptions of a parrot's speech, to present a novel method of dealing with inter-rater reliability which we believe can be applied to situations in which speech from human subjects may be difficult to transcribe. Challenges encountered included (1) a sparse original agreement matrix which yielded an omnibus measure of inter-rater reliability, (2) "lopsided" 2×2 matrices (i.e. subsets) from the overall matrix and (3) categories used by the transcribers which could not be pre-determined. Our novel approach involved calculating reliability on two levels—that of the corpus and that of the above mentioned smaller subsets of data. Specifically, the technique included the "reverse engineering" of categories, the use of a "null" category when one rater observed a behavior and the other did not, and the use of Fisher's Exact Test to calculate r-equivalent for the smaller paired subset comparisons. We hope this technique will be useful to those working in similar situations where speech may be difficult to transcribe, such as with small children.

Keywords Inter-rater reliability · Rater agreement · Fisher's Exact Test · r-Equivalent · Sparse agreement matrix · Speech transcription

Allison B. Kaufman is now in the Department of Ecology and Evolutionary Biology at The University of Connecticut.

A. B. Kaufman (✉)
California State University, San Bernardino, 5500 University Parkway,
San Bernardino, CA 92407, USA
e-mail: akaufman@csusb.edu

E. N. Colbert-White
Department of Psychology, University of Puget Sound, 1500 N. Warner #1046, Tacoma, WA 98416, USA
e-mail: ecolbertwhite@pugetsound.edu

R. Rosenthal
Department of Psychology, University of California, Riverside,
900 University Ave., Riverside, CA 92521, USA
e-mail: robert.rosenthal@ucr.edu

Published online: 13 July 2013

Table 1 2 × 2 Agreement matrix showing 57 % rater agreement

RATER 2	RATER 1	
	Yes	No
Yes	57	21
No	22	0

The associated *r* value in this example was −.27, which is statistically significant in the *opposite* direction

1 Inter-rater reliability and speech transcriptions

By definition, inter-rater reliability calculates the degree of match between two independent observers witnessing the same event, thus providing a measure of how reliable the researcher is in his or her recordings (Tinsley and Weiss 1975). Traditionally, reliability measurements are made in one of two situations. In the first situation, such as a social worker counting aggressive behaviors at a playground, observers are tasked with watching subjects and selecting behaviors they have witnessed from a list of expected behaviors or behaviors of interest. In the second situation, such as that same social worker scoring the intensity of a fight between two children, observers make ratings on scales to quantify observed behavior. Unfortunately, as much as it would be preferable, calculating inter-rater reliability is not so straightforward for all events and behaviors.

Researchers transcribing language, particularly that of inexperienced speakers like children, must be aware of the inherent difficulty of the transcription process. Typically, reliability between observers is calculated via a symmetrical matrix of the potential options a rater might code. These categories are distinct, mutually exclusive, and definitive. For example, if the categories are Red, Green, Blue and Yellow, raters are responsible for observing and coding from these four choices and only these four choices. However, the speech of a child still in the early stages of language development is decidedly ambiguous (Geert and Dijk 2003); and as any parent can attest, distinguishing a child's words is a skill developed only with much practice. In addition to this, language provides a very unique rating situation as the set of potential "categories" to be transcribed (i.e., "coded") consists of every word or sound raters could possibly identify. In this way, the number of categories is practically infinite. The idea of a theoretically infinite set of categories has not yet been addressed, to our knowledge, in methodological literature. Further, despite its relevance to studies involving human speech, data on inter-rater reliability calculations of speech transcriptions are few (Stockman 2010), and empirical studies of transcriptions demonstrate significant error between coders (for discussion, see Lindsay and O'Connell 1995; Geert and Dijk 2003). For example, in a study aimed specifically at testing inter-rater reliability, Stockman (2010) used percent agreement between raters (which is considerably less stringent and less informative than Cohen's kappa; Cohen 1960), and still found very high levels of *dis*agreement. Specifically, only 57 % of overall agreement was achieved across raters on word boundary location in the spontaneous speech of preschool-aged children.[1] This particularly dismal amount of agreement evidences the difficulty of the task at hand.

We describe here data potentially more complex than even children's language, in the hope that the reliability techniques developed will prove useful in other situations. The

[1] Theoretically, in a case such as this, the 57 % agreement can be dramatically inflated. See Table 1 for an example scenario.

scenario presented involves the vocalizations of an African Grey parrot (*Psittacus erithacus*) by the name of Cosmo. A previous study involved transcriptions of Cosmo's vocalizations by two raters (ECW and ABK; Colbert-White et al. 2011). ECW coded Cosmo's vocalizations from video and ABK coded a portion of those sessions to assess reliability. Issues of word ambiguity due to low audio clarity, as well as the lack of a pre-determined coding scheme as discussed above, and the desire to examine reliability at two different levels the coding matrix, led us to develop a novel methodology for reliability calculations which featured the measurement of reliability at both the level of the overall corpus (i.e. body of text) and smaller subsets of interest. We hope that despite development with a non-human speaker, the techniques presented here will be helpful to those working with human subjects.

2 General and specific calculations of reliability

For the analysis of Cosmo's speech, it was desirable to calculate reliability in two different areas. Primarily, it was important to calculate reliability over the entire corpus for the purposes of the Colbert-White et al. (2011) manuscript. However, we also sought information on the coding reliability for smaller subsets of the corpus. These subsets ranged from individual word occurrences (e.g., *hello*) to groupings of similar words (e.g., all words that occurred at the beginning of a phrase). Reliability data on these smaller subsets would allow us to investigate the specific causes of rater disagreement—for example, if disagreement was higher on non-word sounds more training might be required, or if it was higher on words that began phrases, adjustments to the audio might be made to increase clarity. We will examine each of these two levels of reliability separately, as the characteristics of each dictated the use of different methodologies.

An example coding matrix can be found in Table 2. Coding for reliability was done via one minute intervals in which the rater transcribed every sound vocalized by Cosmo, along with the specific time of the vocalization. As Cosmo tends to speak in diverse phrases, this resulted in reliability matrices that were both large and sparse. In Table 2, the example matrix contains 23 different categories coded, but many occurred only once in the course of the minute (see, for example, the word "come"). In addition, categories for coding could not be specified in advance. These categories were, for all intents and purposes, the words Cosmo uttered and therefore were only designated categories after they had been coded by one or more of the raters.

Traditionally, one of the most often used measures of inter-rater reliability is Cohen's kappa (Cohen 1960). Unfortunately, kappa becomes less informative in situations where the coding matrix is larger than 2×2 and $df > 1$ (as was the case with our data); in these situations, kappa becomes an omnibus statistic and it becomes impossible to determine whether, for example, the raters were equally reliable in all coding options, or whether they were excellent at some coding options and poor at others (Rosenthal 2005; Rosenthal and Rosnow 2008). In some situations, even a focused kappa (i.e., 2×2 with $df = 1$) is less informative than an r value, as only an r value can yield a meaningful coefficient of determination or binomial effect size display (BESD; Rosenthal and Rubin 1982; Rosenthal 2005; Rosenthal and Rosnow 2008). We determined the best way to handle the situation was to use a kappa value for overall reliability at the level of the corpus (see below for reasons), and then to use r values to further examine reliabilities of selected subsets of data.

Table 2 Sample reliability coding matrix

		ABK																						
		a	bye	come	cosmo	DB	dogs	DW	for	go	gonna	good	hello	i	ID	love	MWH	null	NWM	okay	on	walk	we're	you
ECW	a	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	bye	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	come	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	cosmo	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	DB	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	dogs	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	DW	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0
	for	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	go	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	gonna	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
	good	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	hello	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	i	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	ID	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	love	0		0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	MWH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	null	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	NWM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	okay	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
	on	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	walk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
	we're	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
	you	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

To illustrate how the matrix is interpreted, rater ECW coded six occurrences of DW [*DOG WHINE/WHIMPER*], but rater ABK was only in agreement at two of these occurrences (coding the other incidences as MWH [*OTHER ONE-NOTE WHISTLE*]and NWM [*OTHER NON-WHISTLE SOUND*]).

3 Measuring reliability at the corpus level

3.1 Other available methods

Many statistics for determining inter-rater reliability are readily available, however, we found none were appropriate for our data. Brennan and Light (1974) developed a test statistic for inter-rater reliability in cases of raters selecting their own categories. The A′ statistic was based upon rater classification of the behavior of two children into either the same or different categories. The benefit of this approach was that the statistic addressed instances in which raters agreed upon what the behavior was, as well as instances in which raters agreed upon what the behavior was not, thus circumventing the main pitfall of the percent agreement statistic (Kaufman and Rosenthal 2009; see Table 2). As Table 1 demonstrates, percent agreement often overestimates the actual agreement between observers, resulting in high percent agreement and an r statistic that is low or even statistically significant in the opposite direction. Brennan and Light (1974) also assumed the marginal totals to be fixed. This is not necessarily the case with the data presented here, as oftentimes one rater heard a vocalization which was not heard by the other rater.

Another statistic, Yule's Q (Montgomery and Crittenden 1977), was considered because of its common usage when reliability coding matrices are sparse—as in the case of our data. Though a Yule's Q calculation might appear appropriate on the surface, the statistic was originally proposed by Montgomery and Crittenden (1977) in a method in which multiple categories from each rater were reduced to a 2×2 table by grouping perceived subcategories. For example, in a situation in which four instances of a behavior are coded into one category by Rater 1 but coded into three smaller, more distinct categories by Rater 2, Montgomery and Crittenden suggested combining Rater 2's categories to match the larger category established by Rater 1. While this may eliminate the problem of different numbers of categories proposed by the raters, it may also artificially inflate their agreement (as new categories are considered subcategories of the original, and therefore may be erroneously recorded as agreements). In addition, when a matrix larger than 2×2 is created this way, the inter-rater reliability statistic becomes an omnibus statistic and thereby is subject to the same disadvantages as Cohen's kappa.

Scott's π (Scott 1955) and Krippendorff's α (Krippendorff 1978) are popular in content analysis research. However, Scott's π does not account for rater bias in coding; that is, a rater's tendency to prefer to code particular categories over others. This was a salient issue as one rater (ECW) was far more familiar with Cosmo's everyday vocabulary than the other. Krippendorff's α is, unfortunately, a very complex calculation which is not readily available in traditional statistical packages.

We considered two other statistics—Hubert's Γ and the J-Index—which also measure inter-rater agreement when categories are developed by the raters themselves (also uncharacteristic of our data set, see below for more details; Hubert 1977; Popping 1983, 1984). However, neither of these reliability statistics is commonly applied to sparse matrices larger than 2×2 (regardless of the complexity of the dataset).

3.2 Our method

Given the number of imperfect statistics for assessing inter-rater reliability in speech transcriptions at the level of the overall matrix, we elected to use the kappa statistic supplemented with additional measurements. These additional measurements would serve to confirm that the reliability found by kappa was due to reliability spread out fairly evenly across all coding

categories (i.e. words), as opposed to raters doing an exceptional job of coding some categories and a poor job of coding others. We calculated reliability in these subsets via *r*-equivalent (Rosenthal and Rubin 1982).

4 Measuring reliability at the token lsevel

4.1 The nature of the categories

Cosmo's vocalizations presented a unique challenge for assessing inter-rater reliability because, to a naïve transcriber, Cosmo's vocal repertoire theoretically could have contained any word in the English language, or any non-word sound, Cosmo could produce. Logistically, this would have made it impossible to pre-establish categories for coding. In addition, a key focus of the Colbert-White et al. (2011) study was Cosmo's ability to create and use novel vocalizations. Providing the second observer (i.e., ABK) with Cosmo's repertoire to use as a coding scheme would have introduced the potential for a priming bias against coding novel vocalizations as novel. For example, if "box" had been provided on an *a priori* list, ABK, hearing the novel utterance "bach" may have been primed to hear (and thus record) it as "box."

As previously mentioned, this meant that coding categories were essentially "reverse engineered" for each minute. That is to say, the coding scheme for a particular minute consisted of the set of words that had been heard at least once by either of the raters during a particular minute of the video clip. There were two immediate consequences of this. First, a "null" category was incorporated for instances in which one rater recorded a vocalization where the other did not (e.g., in Table 2, ABK, but not ECW, recorded "hello" once). The second consequence of the reverse-engineered coding scheme was the resulting necessity of a new template for the calculations presented below, as an existing appropriate statistical package could not be found. This template was developed in Microsoft Excel and the progression of calculations is shown in Table 3.

4.2 The null category

The "null" category was used when one rater coded a word and the other rater coded nothing. In these cases, the word "null" was placed in the transcription as a place holder for the unheard word. This meant that during analysis the null category, by definition, consisted of disagreements. When these instances were removed from calculations involving subsets of data (see below), *r* values increased by .1–.2. The technique of using a null category is, as far as we know, a novel contribution to the inter-rater reliability literature within the study of language. As we were coding from video tape which could be re-watched, the transcription was held to a higher standard of accuracy—by requiring a match for timestamps as well as vocalizations. As a result, the purpose of the null category was to mark places in which only one rater transcribed a vocalization and to assist in maintaining the temporal integrity of the transcription (refer to the 24th s in Table 4 for an example).

4.3 Calculation of *r*-equivalent

As with human speech, each individual minute of Cosmo's speech contained a large number of tokens, very few of which were repeated within the minute-long reliability segments. For every vocalization coded by the observers, there were $n - 1$ words that the observers agreed

Table 3 Fisher's Exact Tests for words in a minute-long reliability coding segment

	YES/YES	NO/YES	YES/NO	NO/NO	p Value	t Value	t^2	$t^2 + df$	$t^2/(t^2 + df) = r^2$	r
a	3	0	0	41	.000076	4.17	17.36	59.36	0.29	0.54
bye	0	0	1	43	1.00	–	–	–	–	–
come	1	0	0	43	0.023	2.06	4.25	46.25	0.09	0.30
cosmo	1	0	0	43	0.023	2.06	4.25	46.25	0.09	0.30
DB	2	1	1	40	0.0093	2.45	5.98	47.98	0.12	0.35
dogs	1	0	1	42	0.045	1.73	2.99	44.99	0.07	0.26
DW	2	4	0	38	0.016	2.22	4.94	46.94	0.11	0.32
for	3	0	0	41	.000076	4.17	17.36	59.36	0.29	0.54
go	3	0	0	41	.000076	4.17	17.36	59.36	0.29	0.54
gonna	3	0	0	41	.000076	4.17	17.36	59.36	0.29	0.54
good	1	0	0	43	0.023	2.06	4.25	46.25	0.09	0.30
hello	0	0	1	43	1.00	–	–	–	–	–
i	1	0	0	43	0.023	2.06	4.25	46.25	0.09	0.30
ID	0	2	0	42	1.00	–	–	–	–	–
love	1	0	0	43	0.023	2.06	4.25	46.25	0.09	0.30
MWH	1	0	2	41	0.068	1.52	2.31	44.31	0.05	0.23
null	0	2	0	42	1.00	–	–	–	–	–
NWM	1	0	3	40	0.091	1.36	1.84	43.84	0.04	0.21
okay	3	0	0	41	.000076	4.17	17.36	59.36	0.29	0.54
on	1	0	0	43	0.023	2.06	4.25	46.25	0.09	0.30
walk	3	0	0	41	.000076	4.17	17.36	59.36	0.29	0.54
we're	3	0	0	41	.000076	4.17	17.36	59.36	0.29	0.54
you	1	0	0	43	0.023	2.06	4.25	46.25	0.09	0.30

YES and NO comparisons are for ABK and ECW, respectively

Table 4 Sample reliability transcriptions

Timestamp	ABK	ECW
:02	good bye	good ID
	cosmo	cosmo
	i	i
	love	love
	you	you
:08	okay	okay
	dogs	ID
	we're	we're
	gonna	gonna
	go	go
	for	for
	a	a
	walk	walk
:11	okay	okay
	dogs	dogs
	we're	we're
	gonna	gonna
	go	go
	for a	for a
	walk	walk
:15	come	come
	on	on
:17	MWH	MWH
:20	DB	DB
:22	NWM	DW
:24	hello	*
:25	NWM	DB
:26	DB	*
:27	DW	DW
:35	NWM	NWM
:40	NWM	DW
:41	DB	DB
:42	DW	DW
:45	okay	okay
	we're	we're
	gonna	gonna
	go	go
	for	for
	a	a
	walk	walk
:48	MWH	DW
:57	MWH	DW

* null; capital letters denote non-word sounds (see Colbert-White et al. 2011 for full repertoire)

Table 5 2×2 Agreement matrix for the word "for" extracted from Table 2 matrix

ECW	ABK	
	Yes	No
Yes	3	0
No	0	41

$p = .000076$ (one tail), $r^{sample} = 1.00$, r-equivalent $= .54$

Table 6 2×2 Agreement matrix for the word "on" extracted from Table 2 matrix

ECW	ABK	
	Yes	No
Yes	1	0
No	0	43

$p = .23$ (one tail), $r^{sample} = 1.00$, r-equivalent $= .30$

Table 7 Calculation of r-equivalent for words that begin phrases

ECW	ABK		
	Yes	No	
Yes	87	16	103
No	12	2,674	2,686
	99	2,690	2,789
	df = 2,787		

Calculation	Value	Result
Fisher's Exact	2.7859E−134	p Value
Inverse t-test	53.60928363	t Value
t^2	2873.955291	t^2
$t^2 + df$	5660.955291	$t^2 + df$
$t^2/(t^2 + df)$	0.5.7680267	r
sqrt $t^2/(t^2 + df)$	0.712516854	r-eqiuvalent

The values in the boxes are totals. For example, there were 87 instances over the entire corpus where the two raters agreed on what the word at the beginning of a phrase was, and 2,674 instances over the entire corpus when the two raters agreed on what the word at the beginning of a phrase was not

that the coded word was *not*. For example, in a minute-long video clip during which Cosmo vocalized 44 tokens, ABK and ECW agreed that the word "for" was uttered three times. In this way, there were $n - 3$, or 41 instances (out of a total of 44 recorded tokens in the minute) in which the observers agreed that the word "for" was *not* spoken. As a result, any subset of data from within the larger matrix would have created a very "lopsided" 2×2 table which would be inappropriate for χ^2 analysis (see Tables 5, 6 for examples).To circumvent this problem, we elected to calculate r-equivalent for particular pairings. To do so, the probability of a particular pattern of agreements/disagreements between the observers was calculated via Fisher's Exact Test (Rosenthal 2005; Rosenthal and Rosnow 2008). Fisher's Exact Test provides accurate p values for low expected-value 2×2 contingency tables that would not fit the theoretical χ^2 distribution (Fisher 1941; Siegel 1956; Snedecor and Cochran 1989). From this p value, an r-equivalent could be calculated. In a sense, in the context of this methodology, the Fisher's Exact Test was used as a tool or a means to an end, as opposed to simply providing a significance test in and of itself. Tables 7 and 8 illustrate the calculation of r-equivalent for inter-rater reliabilities on words that begin Cosmo's phrases and words that are in the middle of Cosmo's phrases (for example, in the phrase "Cosmo's a good

Table 8 Calculation of r-equivalent for words which are inside phrases

ECW	ABK		
	Yes	No	
Yes	81	10	91
No	19	2,738	2,757
	100	2,748	2,848
	df = 2, 846		
Calculation	Value	Result	
Fisher's Exact	2.2098E−126	*p* Value	
Inverse t-test	50.02896524	t Value	
t^2	2502.897363	t^2	
$t^2 + df$	5348.897363	$t^2 + df$	
$t^2/(t^2 + df)$	0.467927723	r^2	
Sqrt $t^2/(t^2 + df)$	0.684052427	r-eqiuvalent	

bird," *Cosmo* would be categorized as beginning the phrase and *a*, *good*, and *bird* would be categorized as in the phrase).

5 Discussion

We presented here a novel technique for the calculation of inter-rater reliability in both overall corpora and smaller subsets thereof, and in cases where the categories to be coded are not specified *a priori*.

The original data described here consisted of a large corpus for which inter-rater reliability could be computed; however any resulting statistic would be an omnibus statistic. As a result, to get a better idea of the inter-rater agreement on smaller subsets of data, 2 × 2 tables were extracted from within the corpus for analysis. Due to the nature of the data, these 2 × 2 tables were necessarily unbalanced, and when 2 × 2 tables are not balanced, the direct computation of *r* from its definition can give highly misleading results (Rosenthal and Rubin 2003). Because this was the case with the data presented here, we used a different statistic, *r*-equivalent (Rosenthal and Rubin 2003). However, computing an accurate *r*-equivalent requires an accurate *p* value, which often cannot be obtained from a χ^2 in this situation. As a result, a more appropriate *p* value was obtained via Fisher's Exact Test (although it should be noted that there are other ways of obtaining accurate *p* values using such resampling techniques as bootstrapping or jackknifing).

In addition to the novel set of calculations used to obtain the *r* value, the situation presented here was such that the coding categories were not specified in advance and therefore were "reverse engineered" after observations were complete. There is much to be investigated with regard to this technique and how it might impact data analysis. For example, is the .1–.2 decrease in *r* caused by the null category acceptable? Or, would the data be more accurately or usefully represented if the incidences in which a word is transcribed by only one rater (what was to become the null category) were simply removed? Conversely, would it be more practical to avoid the situations all together and deem measuring accuracy in seconds too lofty a goal; making it acceptable to adjust and align the transcripts if, for example, the raters coded the same words but at times off by a second or two?

Though the situation of transcribing parrot speech is a novel one, it is our hope that the method is useful for situations in which (1) post-hoc categories must be developed for inter-rater reliability, (2) a null category is useful, (3) the overall matrix will yield a less helpful omnibus statistic, and/or (4) subsets of the overall matrix are unbalanced enough to preclude the use of a χ^2 test. We hope that the techniques presented here will be of use to researchers studying speech and language, and we look forward to further discussion on the validity and implications of our method for data analysis.

References

Brennan, R.L., Light, R.J.: Measuring agreement when two observers classify people into categories not defined in advance. Br. J. Math. Stat. Psychol. **27**, 154–163 (1974)

Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**, 37–46 (1960)

Colbert-White, E.N., Covington, M.A., Fragaszy, D.M.: Social context influences the vocalizations of a home-raised African grey parrot (*Psittacus erithacus erithacus*). J. Comp. Psychol. **125**, 175–184 (2011). doi:10.1037/a0022097

Fisher, R.A.: Statistical methods for research workers. Oliver & Boyd, Edinburgh (1941)

Van Geert, P., Van Dijk, M.: Ambiguity in child language: the problem of interobserver reliability in ambiguous observation data. First Lang. **23**, 259–284 (2003)

Hubert, L.: Nominal scale response agreement as a generalized correlation. Br. J. Math. Stat. Psychol. **30**, 98–103 (1977)

Kaufman, A.B., Rosenthal, R.: Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. Anim. Behav. **78**, 1487–1491 (2009)

Krippendorff, K.: Reliability of binary attribute data. Biometrics **34**, 142–144 (1978)

Lindsay, J., O'Connell, D.C.: How do transcribers deal with audio recordings of spoken discourse? J. Psycholinguist. Res. **24**, 101–115 (1995)

Montgomery, A.C., Crittenden, K.S.: Improving coding reliability for open-ended questions. Public Opin. Q. **41**, 235–243 (1977)

Popping, R.: Traces of agreement: on the DOT-product as a coefficient of agreement. Qual. Quant. **17**, 1–18 (1983)

Popping, R.: Traces of agreement: on some agreement indices for open-ended questions. Qual. Quant. **18**, 147–158 (1984)

Rosenthal, R.: Conducting judgment studies: some methodological issues. In: Harrigan, J., Rosenthal, R., Scherer, K. (eds.) The new handbook of methods in nonverbal behavior research, pp. 199–236. Oxford University Press, New York (2005)

Rosenthal, R., Rubin, D.B.: A simple, general purpose display of magnitude of experimental effect. J. Educ. Psychol. **74**, 166–169 (1982)

Rosenthal, R., Rubin, D.B.: r-equivalent: a simple effect size indicator. Psychol. Methods **8**, 492–496 (2003)

Rosenthal, R., Rosnow, R.: Essentials of behavioral research: methods and data analysis. McGraw-Hill, New York (2008)

Scott, W.: Reliability of content analysis: the case of nominal scale coding. Public Opin. Q. **17**, 321–325 (1955)

Siegel, S.: Nonparametric statistics for the behavioral sciences. McGraw-Hill, New York (1956)

Snedecor, G.W., Cochran, W.G.: Statistical methods. Iowa State University Press, Ames (1989)

Stockman, I.: Listener reliability in assigning utterance boundaries in children's spontaneous speech. Appl. Psycholinguist. **31**, 363–395 (2010)

Tinsley, H.E.A., Weiss, D.J.: Interrater reliability and agreement of subjective judgments. J. Couns. Psychol. **22**, 358–376 (1975)