Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

STATISTICAL MODELS FOR ASSESSING AGREEMENT IN METHOD COMPARISON STUDIES WITH HETEROGENEOUS RANDOM RATERS AND REPLICATE MEASUREMENTS

By Claus Thorn Ekstrøm* and Bendix Carstensen

University of Copenhagen and Steno Diabetes Center

Agreement between methods for quantitative measurements are typically assessed by computing limits of agreement between pairs of methods and/or by illustration through Bland-Altman plots. We consider the situation where the observed measurement methods are considered a random sample from a population of possible methods, and discuss how the underlying linear mixed effects model can be extended to this situation. This is relevant when, for example, the methods represent raters/judges that are used to score specific individuals or items. In the case of random methods, we are not interested in estimates pertaining to the specific methods, but are instead interested in quantifying the variation between the methods actually involved making measurements, and accommodating this as an extra source of variation when generalizing to the clinical performance of a method. In the model we allow raters to have individual precision/skill and permit linked replicates (*i.e.*, when the numbering, labeling or ordering of the replicates within items is important). Applications involving estimation of the limits of agreement for two datasets are shown: A dataset of spatial perception among a group of students as well as a dataset on consumer preference of French chocolate. The models are implemented in the `MethComp` package for R.

1. Introduction. Comparison of methods for quantitative data is concerned with how well two methods agree on the measurement of an item. The interest is not on testing a hypothesis that the mean of the the two methods are identical, but on estimating the size and the standard deviation of the difference between them.

Several authors have considered statistical models for assessing agreement with continuous measurements (Barnhart, Haber and Lin, 2007; Rousson, Gasser and Seifert, 2002; Bland and Altman, 2007). Carstensen, Simpson and Gurrin (2008) considered a mixed effect model for computing the prediction limits (often termed the *limits of agreement*) for designs with repli-

*To whom correspondance should be addressed.

Keywords and phrases: Agreement, Limits of agreement, Methods comparison, Mixed models, Random raters

cate measurements on each item. This approach provides a full statistical model that can compute the limits of agreement in the presence of replicate measurements instead of relying on summary measures. Moreover, it forces the investigator to focus on the nature of replicate measurements: are they exchangeable within each method by item stratum or only within items (*i.e.*, the replicates are linked).

Here we consider the situation where we are not interested in determining the bias and agreement between two *specific* methods but are interested in the agreement between two random methods. Thus, the methods included in the study are regarded as a random sample of possible methods from a larger population of available methods; random methods are relevant, for example, when a group of judges/raters are asked to rate a set of items using a predetermined scale. This is a common occurrence for example when medical doctors are asked to give second or third opinions on measurements taken on patients. We are interested in how well medical doctors in general agree on the scoring of a particular condition, more so than determining how well specific doctors compare to each other. However, the agreement among random raters are not restricted to humans as measurement measurement methods. A similar approach can be used in combination with, say, machine learning where a large number of (random) regression trees are generated and we wish to compare how well two randomly selected regression trees predict the same value. Here the regression trees act as raters and we wish to investigate how well two randomly chosen trees agree.

The focus on random raters (as opposed to fixed methods/raters) is not new and is a often found in relation to the calculation of intraclass correlation coefficients (ICC) for random factorial design (Shoukri, 2010; Gwet, 2012). However, they are typically concerned with estimating the ratio of the intraclass variation to the total variance. Here, our focus is on limits of agreement and we present a model that easily accommodates individual precision/skill for each rater and handles unbalanced designs where raters may rate different subsets of items.

Limits of agreement estimated for random raters mimics the reproducibility coefficient (ISO, 1994). The reproducibility coefficient includes the variability among different laboratories, when observations are carried out under reprocibility conditions at different laboratories. In that sense, the reproducibility coefficient also seeks to increase the variability by allowing the laboratories to be different, but again we wish to extend that concept to situations with individual skill levels and with unbalanced designs.

Section 2 sets up the necessary mixed effect model that accommodates random raters, and Section 3 extends the model to the situation where repli-

cate measurements from the methods are available on each item. Section 4 discusses repeatability while the model is applied to two datasets with random raters in Section 5.

2. Models for agreement among random raters. The traditional setup for comparison of measurements methods is one where exactly one measurement is taken with each method on each item (here, 'item' can refer to, for example, an individual, a sample, or an image). The limits of agreement for two methods are computed as the prediction interval for the difference between future measurements taken by the two methods on a new item or, equivalently as a prediction interval for a measurement on an item by method B, given a measurement by method A and *vice versa*. The focus of such studies is typically the comparison (and possibly prediction) between a few specific methods of interest.

When multiple raters are compared, the raters play the role of methods, but we are not necessarily interested in the difference between any two specific raters. Instead we consider each rater as a "random" rater/judge/expert from an (essentially infinite) population of raters. Thus, when we compute the prediction limits for the difference between measurements by two randomly chosen raters we should ensure that the variation between the random raters is taken into account in the modeling and that the variation is included in the computation of the prediction interval. The situation with random raters arises for example in medical situations where a medical doctor is asked to give a second opinion on the measurement from a patient.

If we carry over the two-way analysis of variance model from the traditional Bland-Altman setup we model the value for a measurement taken on item i by rater m as:

$$\begin{aligned} Y_{mi} &= \mu_i + b_m + e_{mi}, \\ b_m &\sim N(0, \xi^2), \\ e_{mi} &\sim N(0, \sigma^2), \end{aligned} \tag{1}$$

where μ_i is the "true" value for item i, b_m is a random effect that models a random bias for measurements taken by rater m, ξ is the variation of biases between raters, and σ is the individual variation of rater m (Bland and Altman, 1999; Carstensen, Simpson and Gurrin, 2008). Note that above we have implicitly assumed that the individual variation for each rater is the same — we will relax this assumption below.

The "true" value for item i is of course arbitrary; a constant may be added to all μ_is provided it is subtracted from the b_ms; but since we have constrained the mean of the b_ms to be 0, the μ_is represent the average assess-

ment of item i by the set of raters at hand. Note that since the raters (or more precisely, the variation between them) is the focus of interest, the μ_is are essentially nuisance parameters. Even if only one measurement per rater and item is available it is possible to estimate all parameters in model (1), so in order to produce limits of agreement for random raters in this simple setup we just need to estimate the mean rater-specific variation across raters.

The limits of agreement for the difference between measurements taken by two random raters (m and m') on the same item corresponds to the prediction interval of the difference $Y_{mi} - Y_{m'i}$. If measurements by different raters are assumed to be independent we get that the limits of agreement (prediction interval) for the difference between the two raters on a new item is

$$0 \pm z\sqrt{\text{Var}(Y_{mi} - Y_{m'i})} = 0 \pm z\sqrt{\text{Var}(Y_{mi}) + \text{Var}(Y_{m'i})}, \tag{2}$$

where z is the quantile corresponding to the desired level of the prediction interval. Under model (1), the 95% limits of agreement simplifies to

$$0 \pm 1.96 \times \sqrt{2 \times (\xi^2 + \sigma^2)} \approx 0 \pm 2.8\sqrt{\xi^2 + \sigma^2}. \tag{3}$$

In practice we will replace the parameters in (3) by their corresponding estimates. The value 1.96 can of course be replaced by a suitable quantile from the t distribution; the normal convention is to use 2 (which incidentally is the 97.5% quantile of the t distribution with 60 degrees of freedom). In the rest of this paper we shall use 2 for convenience.

Model (1) can be extended to allow for raters to have different skill/precision such that some raters can be very precise (*i.e.*, have low residual variation) while others can be less precise (*i.e.*, have high residual variation). The value for a measurement taken on item i by rater m becomes

$$\begin{aligned} Y_{mi} &= \mu_i + b_m + e_{mi}, \\ b_m &\sim N(0, \xi^2), \\ e_{mi} &\sim N(0, \sigma_m^2), \end{aligned} \tag{4}$$

where σ_m is the individual variation of rater m. Allowing for heterogeneity among raters extends the traditional modeling and the variance for rater m on item i becomes

$$\text{Var}(Y_{mi}) \quad = \quad \xi^2 + \sigma_m^2$$

where we let the individual residual variances, σ_m^2, follow some distribution of the residual variances which has support on the positive real numbers.

Thus, the distribution of the residual variances is central for the prediction of differences between raters; it represents the distribution of "skill" among available raters. If prior knowledge about the precision distribution is available then that information can be used to model the distribution of the residual variances. In the following we assume that we have no prior knowledge about the precision and/or that we have too little information to be able to verify any distributional assumptions about the residual variances that we might have.

The law of total variance provides the variance of a measurement from a *randomly chosen rater* from the population when applied to a fixed item:

$$\begin{aligned} \mathrm{Var}(Y_{mi}) &= \mathrm{Var}_\sigma\big(\mathrm{E}(Y_{mi}|\sigma_m^2)\big) + \mathrm{E}_\sigma\big(\mathrm{Var}(Y_{mi}|\sigma_m^2)\big) \\ &= \mathrm{Var}_\sigma(\mu_i) + \mathrm{E}_\sigma(\xi^2 + \sigma_m^2) \\ &= \xi^2 + \mathrm{E}_\sigma(\sigma_m^2). \end{aligned} \tag{5}$$

E_σ and Var_σ represent means and variances based on the underlying distribution of the residual variances. We can estimate the variance for a randomly chosen rater using the empirical counterpart of (5):

$$\widehat{\xi^2} + \widehat{\mathrm{E}(\sigma_m^2)} \approx \widehat{\xi^2} + \frac{1}{M}\sum_{m=1}^{M} \widehat{\sigma_m^2} \tag{6}$$

where M is the number of raters available in the dataset. Hence, the estimate for the 95% limits of agreement between two randomly chosen raters becomes

$$0 \pm 2 \times \sqrt{2 \times \big(\widehat{\xi^2} + \frac{1}{M}\sum_{m=1}^{M} \widehat{\sigma_m^2}\big)} \tag{7}$$

Equation 5 states that the variance of a measurement by a randomly chosen rater is the sum of between rater variance and the average within-rater variance. The estimate relies heavily on the assumption that the sample of raters at hand is representative of the population of future raters. Moreover, the within-rater variation might be poorly determined if the number of raters in the study is small.

Even without replicate measurements on each item it is possible to estimate the individual residual variance for each rater as long as each rater has scored at least two items, because model (1) by the very nature of the randomness of raters must impose an assumption of 0 average difference between raters. This means that the estimate of the single rater's variation is strongly dependent on the other raters' results, because it is essentially

	I					II					III				
	Rater					Rater					Rater				
Item	*A*	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*
1	•	•	•	•	•	•	•				•	•	•	•	
2	•	•	•	•	•	•	•				•	•	•	•	
3	•	•	•	•	•	•		•			•	•	•	•	
4	•	•	•	•	•	•		•			•	•	•	•	
5	•	•	•	•	•	•			•		•	•	•		•
6	•	•	•	•	•	•			•		•	•	•		•
7	•	•	•	•	•	•				•	•	•	•		•
8	•	•	•	•	•	•				•	•	•	•		•

FIG 1. *Examples of designs with 8 items and 5 random raters. A dot indicates that an item was scored by the rater. I) Balanced design where each rater scores every item. II) "Teacher design" where a single expert, A, scores every item while trainees (raters B–E) each score a few items. III) "Chief phycisians design" where a large number of medical students (raters A–C) examines every item/patient while a smaller group of chief phycisians (raters D–E) oversee the students as they do the rounds.*

the variation around the common means (μ_i). It also emphasizes the crucial nature of the assumptions behind *random* raters, a truly random sample of raters is required to ensure that the empirical means across the raters at hand is a fair approximation of the true population mean. To put it another way, the generalizations are only valid for a population of raters of which those in the sample at hand can be considered a random sample. And predictions of precision for future ratings also require that the sample of items for which the prediction is made is also a random sample of the population of items.

The total variance for a random rater in the situation with homogeneous individual variances, $\xi^2+\sigma^2$, resembles the variance for a random rater in the situation with heterogeneous individuals variation, $\xi^2+\mathrm{E}_\sigma(\sigma_m^2)$. However, we cannot generally use σ^2 as replacement for the mean individual rater variance $\mathrm{E}_\sigma(\sigma_m^2)$ and then just use the simpler homogeneous model, (1), to estimate the variance components. If the design is balanced, *i.e.*, all raters have scored the same number of items, then we can use (1) even in the heterogeneous case. If the design is unbalanced then it is necessary to estimate $\mathrm{E}_\sigma(\sigma_m^2)$ by $\frac{1}{M}\sum_{m=1}^{M}\widehat{\sigma_m^2}$ since otherwise we give increased weight to raters that have scored more items (see Figure 1 for examples). Model (1) can only replace (4) if each rater has scored exactly the same number of items.

3. Models for agreement among random raters with multiple measurements. It is not uncommon to have situations where multiple

measurements by each rater on some items are available (Bland and Altman, 2007). Multiple measurements for combinations of raters and items makes it possible to estimate the repeatability of a random rater and provides better estimates of the individual residual variations σ_m^2.

As mentioned above, even if raters are measuring each item only once, we can still estimate the residual variation of each rater from model (4); it will simply be the variation around the common item-means. This is where the situation with random raters differs from comparing specific measurement methods. Because raters are considered random, we must necessarily assume that the mean difference between two raters is 0, and by that token that any deviation from 0 is random. This means that the implicit assumption of randomly chosen raters is heavily exploited in the case without replicate measurements.

To assess the variability of the *precision* of a random rater it is mandatory to have multiple measurements of each rater for some item(s); in that case observations are classified by replicate too, so we need a more elaborate model.

If multiple measurements by each rater on an item exist then we can use the following extension of the model (4) for the rth measurement by rater m on item i:

$$\begin{aligned} Y_{mir} &= \mu_i + b_m + a_{ir} + c_{mi} + e_{mir}, \\ b_m &\sim N(0, \xi^2) \\ a_{ir} &\sim N(0, \omega^2) \\ c_{mi} &\sim N(0, \tau_m^2) \\ e_{mir} &\sim N(0, \sigma_m^2) \end{aligned} \tag{8}$$

Note that two new variance components have been added relative to the simple model in Equation 4 and that μ_i is still fixed since agreement is concerned with the variation in scores of one specific new item. The first variance component, ω, is the variation between replication instances; as such it is in principle irrelevant for the comparison of raters. The second, τ_m, is the variation between items within each method — a rater-specific interaction with the items. It represents the variability of a rater across items; that is how a specific rater's measurements varies relative to the *average* measurement by all raters on a given item. Thus, this is a variance component whose size for the individual rater is very strongly tied to the concept of randomly chosen raters, in the sense that the estimates of τ_m will depend on the sample of raters to a much larger degree than will the estimates of σ_m which are entirely

estimates of the individual rater's variation around his own measurement mean.

For situations with multiple measurements by each rater we can follow the same arguments that led to (3) and get limits of agreement between two randomly chosen raters as:

$$0 \pm 2 \times \sqrt{2 \times \left(\hat{\xi}^2 + \frac{1}{M} \sum_{m=1}^{M} (\hat{\tau}_m^2 + \hat{\sigma}_m^2) \right)}. \tag{9}$$

where we have taken the average item-by-rater variation between raters into account. Note that the measurements by different methods are no longer independent because of the term a_{ir} but these terms cancel out when computing $\mathrm{Var}(Y_{mir} - Y_{m'ir})$, because indices i and r are identical for the two terms.

However, repeated measurements come in two guises: exchangeable and linked (Bland and Altman, 2007; Carstensen, Simpson and Gurrin, 2008). Exchangeable observations arise if the order of the observations for a given combination of rater and item is irrelevant. In model (4) this corresponds to having $\omega^2 = 0$ since there is no common variation within items from replicate to replicate.

Replicates are *linked* if the first replicate by all raters are made at the same time (or under similar circumstances), and if the second replicate by all raters are made at the same time too, etc. This means that the *numbering* (but not necessarily the *ordering*) of the replicates carries some information about similar circumstance of the measurement. Linked replicates occur, for example, when the measurements are taken over a longer period of time, or if there is some inherent "memory" in the raters (*i.e.*, they can remember earlier scores on the same item). Linked repeated observations correspond to having $\omega^2 > 0$.

Formula (9) is used to estimate limits of agreement for both exchangeable and linked repeated observations. However, for exchangeable repeated observations we fit model (4) without the variance component ω^2 and thus the two setups might result in different limits of agreement despite using the same formula.

4. Repeatability. The limits of agreement are not always the only issue of interest — the assessment of repeatability and reproducibility for either a specific or randomly chosen rater may be of interest in their own right. In particular, if a rater has large variation between replicates on the same item then the repeatability and agreement with other raters will be poor.

Repeatability can only be assessed when multiple measurements on the same item by each rater are available.

In typical assessment of *specific* raters, the repeatability coefficient for a method is defined as the upper limit of a prediction interval for the absolute difference between two measurements by the same rater on the same item under identical circumstances.

When replicates are exchangeable, the difference between two replicate measurements, r and r' taken by rater m on item i is

$$Y_{mir} - Y_{mir'} = e_{mir} - e_{mir'} \tag{10}$$

so the repeatability is based only on the residual standard deviation, *i.e.*, $2.8\sigma_m$. For linked replicates this difference becomes

$$Y_{mir} - Y_{mir'} = a_{ir} - a_{ir'} + e_{mir} - e_{mir'}, \tag{11}$$

and the variation between replicates taken on the same item should be factored into the calculation of the repeatability coefficient which then becomes $2.8\sqrt{\omega^2 + \sigma_m^2}$. Note that these two results are conditional on rater m since we considered specific raters.

In the case of *random* raters, the repeatability coefficient has a slightly different meaning because we cannot hinge it on estimates of variances from any specific rater — they will just be a random set of variances. Instead the repeatability must refer to the average/expected repeatability, or even, if we cling to the traditional definition, the variability of the repeatability as we may expect to see it in a sample of raters.

Thus for exchangeable replicates we get that the repeatability is

$$E_\sigma(2.8\sigma_m)$$

which is estimated by

$$2.8\frac{1}{M}\sum_{m=1}^{M}\widehat{\sigma_m}, \tag{12}$$

when raters are allowed to have different variances and as 2.8σ, otherwise. Likewise, for linked replicates we get

$$2.8\frac{1}{M}\sum_{m=1}^{M}\sqrt{\widehat{\omega^2} + \widehat{\sigma_m^2}}. \tag{13}$$

The latter argument assumes that the variability between replication occasions can be considered representative of future scenarios. If the replicates

are taken under substantially different circumstances, then the variance component ω may be considered irrelevant for the repeatability and the repeatability coefficient should be based on the measurement errors alone, *i.e.*, use $2.8\frac{1}{M}\sum_{m=1}^{M}\widehat{\sigma_m}$. However, if indeed the replicates are taken under substantially different circumstances it may be argued that we are not really measuring the same item repeatedly. Instead the effects of differing replication circumstances could be modeled by a systematic effect. Hence there is no subject-matter-free way of defining repeatability from the variance components in the models.

5. Example applicaitons.

5.1. *Spatial perception of point swarms.* Some people have keen spatial perception and are able to almost instantaneously give a reasonable guess of, for example, the number of individuals in a crowd. We collected data on spatial perception from participants attending a course on comparison of measurement methods. Thus, we will use attendees of the course as a random selection of raters and will try to determine how well two random raters agree on assessing the number of points in a point swarm.

Ten pictures of scattered points were generated, and the purpose of the exercise was to estimate the number of points in each picture after having viewed the picture for 5 seconds. The number of points in each picture were between 24 and 120. Each picture were shown to the 17 raters three times (rotated and in random order to prevent recollection of the pictures), without telling them about the replication structure. The dataset is available as the dataset `Ancona` (collected at a pre-conference course at the VIth conference of Società Italiana di Statistica Medica ed Epidemiologia Clinica (SISMEC) held in September 2011 in Ancona, Italy) found in the `MethComp` package for R.

We can get an overview of the data by plotting scores of the raters (and the replicates within raters) for each item (picture) as shown in Figure 2, separately for original scale and log-transformed data. The plots in Figure 2 provided pretty clear indications that a log-transform might provide a better fit to the data, since the variances are increasing for the analysis on the original scale, but largely stable for the log-transformed.

Using formula (9) we get that the limits of agreement between two random raters is ± 68.02, thus there is a 95% probability that two randomly selected raters differ less than 68 points in their assessment — not impressive since the range of values are from 24 to 120.

The limits of agreement for the log-transformed data is 1.01. Since this estimate is based on the standard deviations of the log-transformed data

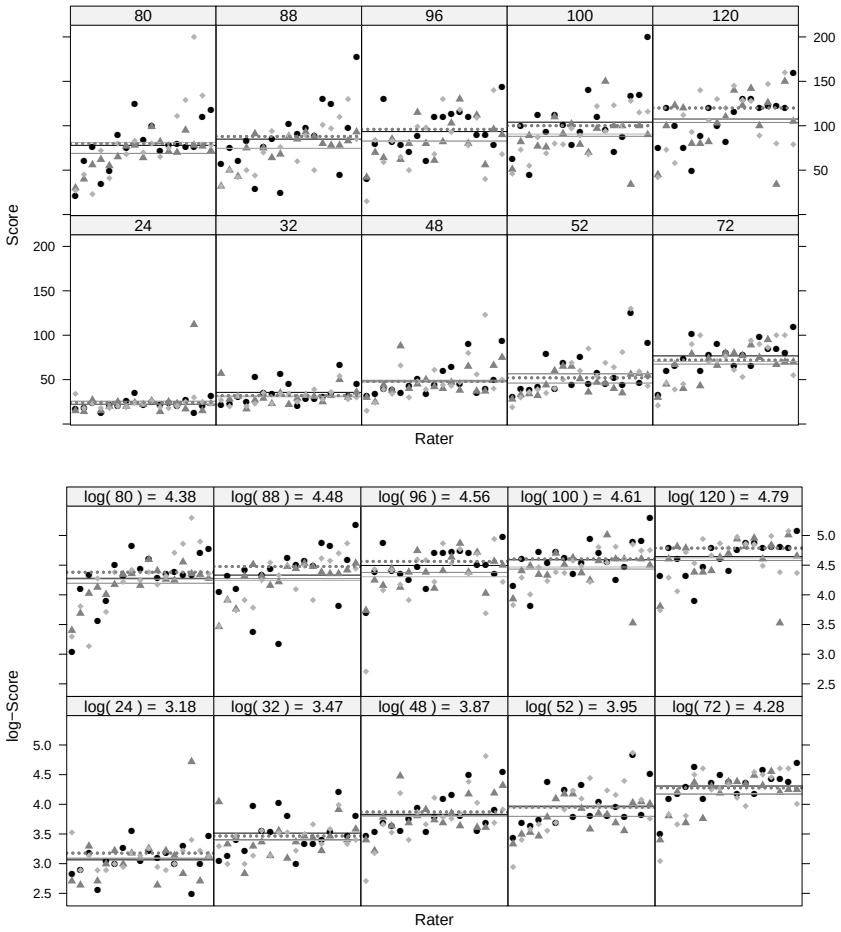

FIG 2. *Illustration of the Ancona data. Each panel shows a specific item and the columns of dots within each panel represent the scores given by each rater (with the true number of points listed above each panel). The symbols/colouring corresponds to replicates (with circles, triangles, and diamonds corresponding to first to third replicate, respectively). Raters are listed in the same order within all panels and are sorted according to their average log-score across all items and replicates. The upper graph shows counts on the original scale, the lower after log-transform. It appears that variances are reasonably homogeneous in the lower panels, but not in the upper. The solid horizontal lines correspond to the mean score within replicates, the dashed line to the true number of points.*

TABLE 1
Individual repeatability coefficients ($2.8\sigma_m^2$) for the 17 different raters.

33.9	29.5	46.6
30.9	48.0	43.8
57.7	36.2	122.4
34.8	52.9	53.7
52.4	23.6	88.0
46.2	41.5	

the estimated variance components essentially represent coefficients of variation, see *e.g.*, Carstensen (2010, chapter 9). To convert this to a sensible number applicable on the original (count) scale, we take the exponential function of these values to get a multiplicative factor. The limits of agreement on the log scale is ± 1.01, so the ratio between the score from two raters (largest/smallest) is with 95% probability less than $\exp(1.01) = 2.75$ for assessments of the number of points the same picture.

We can illustrate the relationship between the two transformations (identity or log) by making a graph for converting between two random raters. The mean conversion line will necessarily be the identity line, so the plot will basically show envelopes of where we can expect to find observations. We add points from random pairs of raters so we can compare the estimated envelopes to the observed data.

Also note that the envelope based on the log-transformation is with straight lines because we have a model where the conversion between methods (raters) is forced to go through 0. The result is shown in Figure 3, where it is clear, that the analysis based on the log-transformed data (limits shown with the dashed lines in Figure 3) captures the variation and differences among raters much better than the untransformed data (the solid lines) for all count levels.

5.1.1. *Repeatability.* The repeatability coefficient for rater m is calculated as $2 \times \sqrt{2\sigma_m^2} \approx 2.8\sigma_m$ provided we assume exchangeability for the replicates. For the Ancona data the individual estimated repeatability coefficients on the original scale are the 17 values shown in Table 1 and the mean repeatability coefficient is 49.53. Broadly speaking we can say that the repeatability of "academic point-counters" (such as represented by the course participants at the SISMEC conference in Ancona) is 49.53 points, meaning that assessing the same picture twice by the same randomly chosen rater will produce two guesses closer than about 50 with probability 95%.

If replicates are linked we should include the between-replicates variation in the calculation of the repeatability, if this variation is considered

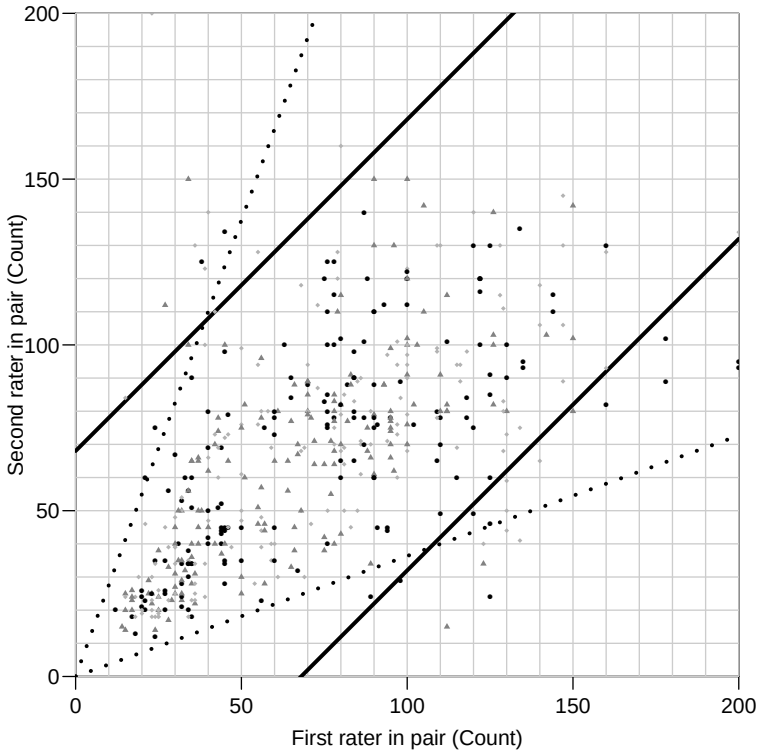

FIG 3. *Visualization of the LoA from the Ancona data. The solid lines are from a model using original count data, the dashed lines from a model for log-transformed data. The points are from 16 randomly chosen pairs of raters.*

unavoidable (and universal) between replicates. In the case of the Ancona experiment, it is debatable whether replicates consisting of showing a rotated version of the picture are replicates in the repeatability sense. But if this is considered so, then we should calculate the individual repeatability coefficients using (11) and compute the mean repeatability coefficient among raters which then becomes 50.23. The mean repeatability is virtually the same for the Ancona dataset whether we consider the replicates as exchangeable or as linked.

We saw that there was some indication that data was better described by a constant variance model for the log-transformed counts. In that case we should compute the repeatability on the multiplicative scale, *i.e.* we should basically do the same calculations on the log-transformed data. Again, these are ("expected") upper limits of absolute differences of natural-log transformed data, so not readily interpretable. If we take the exponential of these, we will get upper limit of a 95% prediction interval for the *ratio* of the larger to the smaller rating between two replicates. Hence, assessing the same picture twice by the same rater will on average produce guesses that are within a factor 2.08 of each other with probability 95%.

All computations shown here are available as a detailed R script with additional comments at
`www.biostatistics.dk/agreement/ancona.pdf`.

5.2. *Consumer evaluation of chocolate.* Sensometrics is typically concerned with quantifying differences between a set of brands or products. The brands are typically scored on several features simultaneously, and the scores are aggregated or analyzed with a multivariate model. If we focus on just a single feature then the products or brands are the items and the consumers that rate the products are the raters. Note that the objective is different from traditional sensometrics: We are interested in evaluating how two randomly picked consumers rate the same product.

The data used here refer to six varieties of chocolates sold in France. Each chocolate was evaluated on a structured scale from 0 to 10, by 222 consumers, according to their liking (0) or disliking (10) (Husson, Le and Cadoret, 2014). For our purpose the six varieties of chocolate will act as items and the 222 consumers as raters and we treat the score as a continuous outcome. Each consumer has only rated each chocolate brand once so there are no replicates. The data can be found in the `SensoMineR` package for R.

An overview of the data can be seen from a violin plot of the scores for each item (chocolate brand) as shown in Figure 4. The densities in Figure 4 suggest that the consumers have very varying opinion on the chocolate scores

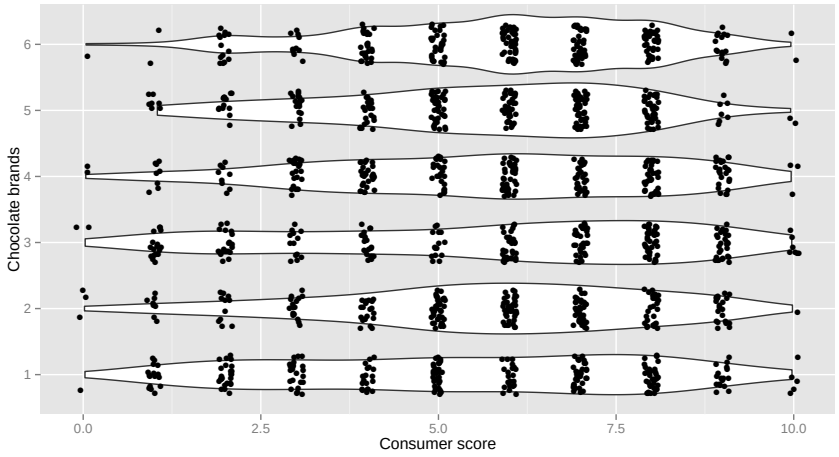

FIG 4. *Illustration of the chocolate data. Each line shows a violin plot and the actual scores for a particular chocolate brand. The scores are shown with a slight jitter along the x axis which is why scores below 0 and above 10 appear in the plot.*

for all the brands, and that there are no clear substantial difference in the brand scores. The plot also indicates that it may be problematic to get to random consumers to agree on the chocolate quality for any brand.

The limits of agreement between two random consumers is 6.66. In other words there is a 95% probability that two randomly selected consumers differ less than 6.7 points in their assessment of the score of a chocolate. Since the scores range from zero to ten the limits of agreement must be said to be very wide.

Figure 5 shows the estimated individual residual standard deviation for each rater and it is clear that there is a substantial heterogeneity in the individual variances. Essentially we could compute the repeatability using (12) if we assume exchangeability, but since we have no actual replicates of measurements on the same item in this dataset we refrain from computing it.

6. Discussion. In this paper we have extended the modeling approach of Carstensen, Simpson and Gurrin (2008) to address the question of precision when the methods represent a random sample of observers (raters) that are asked to produce a measurement. In a sense this is analogous to the binary case that was considered in Fleiss, Levin and Paik (2004). Our extension incorporates the methods as random effects in the model and in that vein, the prediction of the difference between measurements by two random

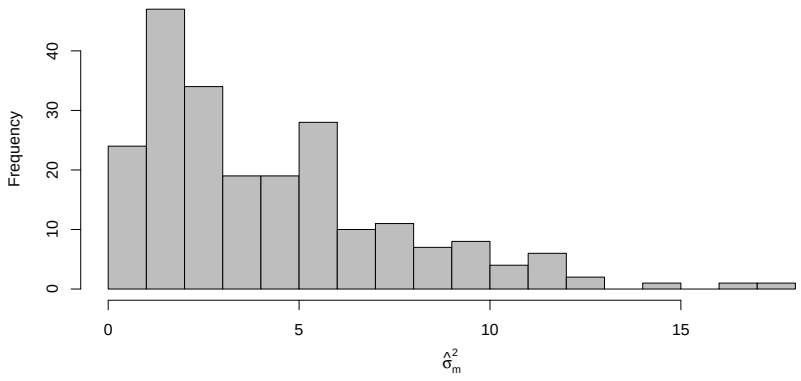

FIG 5. *Individual repeatibility coefficients for the chocolate data.*

raters on a future item will inevitably be a prediction that has zero mean, and where only the variation between and within raters is involved. The proposed modeling approach can handle individual skills of the raters as well as it accommodates designs where not every rater necessarily needs to assess every item.

The calculation of the limits of agreement relies on the law of total variance so it should be noted that the approximations used may be rather crude when the number of available raters is small. It should be emphasized that predictions of differences between future randomly chosen raters is highly dependent on the assumption that the sample at hand is actually representative and that there is no way to test this assumption.

Choudhary (2008) presents the methodology for computing limits of agreement based on tolerance intervals instead of the traditional prediction intervals that we also use here. Tolerance intervals are more desirable for limits of agreement than prediction intervals because we can attach a confidence level to the prediction interval so we have better control over the actual coverage level instead of just an average asymptotic level. Bland and Altman (1999) also suggest to use tolerance limits for limits of agreement with small sample sizes. The prediction and tolerance intervals will be identical when the number of observations increases so the benefit of the tolerance interval is largest for small samples. There are two reasons that we do not consider tolerance intervals in this manuscript. First we need large samples anyway to obtain useful estimates of the variance components in model (8). Second, in order

to obtain tolerance intervals — even for small samples — a double bootstrap approach is needed for the computations. Unless the experimental design is balanced a lot of extra assumptions on resampling units and exchangeability is needed for the bootstrap approach to be valid. Tolerance intervals are preferred, but we use the prediction intervals since we have already made the implicit assumption that we have a sufficiently large sample for the estimates to be useful through prediction intervals.

References.

Barnhart, H. X., Haber, M. J. and Lin, L. (2007). An overview of assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics* **17** 529–569.

Bland, J. M. and Altman, D. G. (1999). Measuring Agreement in Method Comparison Studies. *Statistical Methods in Medical Research* **8** 135–160.

Bland, J. M. and Altman, D. G. (2007). Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics* **17** 571–582.

Carstensen, B. (2010). *Comparing Clinical Measurement Methods: A Practical Guide*. Wiley.

Carstensen, B., Simpson, J. and Gurrin, L. C. (2008). Statistical Models for Assessing Agreement in Method Comparison Studies with Replicates Measurements. *The International Journal of Biostatistics* **4** Article 16.

Carstensen, B., Gurrin, L., Ekstrøm, C. T. and Figurski, M. (2013). MethComp: Functions for analysis of agreement in method comparison studies. R package version 1.22.

Choudhary, P. K. (2008). A tolerance omterval approach for assessment of agreement in method comparison studies with repeated measurements. *Journal of Statistical Planning and Inference* **138** 1102–1115.

Fleiss, J. L., Levin, B. and Paik, M. C. (2004). *Statistical Methods for Rates and Proportions*, 3rd ed. The Measurement of Interrater Agreement. John Wiley & Sons, Hoboken, NJ, USA.

Gwet, K. L. (2012). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters*, 3rd ed. Advanced Analytics.

Husson, F., Le, S. and Cadoret, M. (2014). `SensoMineR`: Sensory data analysis with R R package version 1.20.

ISO (1994). Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions Technical Report, International Organization for Standardization.

Rousson, V., Gasser, T. and Seifert, B. (2002). Assessing interrater, interrater and test-retest reliability of continuous measurements. *Journal of Statistical Planning and Inference* **21** 3431–3446.

Shoukri, M. M. (2010). *Measures of Interobserver Agreement and Reliability*, 2nd ed. Chapman & Hall.

R Core Team (2012). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0.

Biostatistics, Department of Public Health,
Øster Farimagsgade 5B, 1014 Copenhagen, Denmark.
E-mail: ekstrom@sund.ku.dk

Steno Diabetes Center,
Niels Steensens Vej 2, 2820 Gentofte.
E-mail: bxc@steno.dk