Measuring and Optimizing the Quality of Crowd-sourced Human Scoring

Gary Feng, Lei Chen

November 16, 2015

Contents

1 Introduction

The goal of this paper is to explore ways in which we can simultaneously increase the quality of human ratings of multimodal performances while reducing the cost and time required. The scoring approach we take is to use multiple minimally trained human raters, and to rely on the average (or other forms of aggregated) rating as a more reliable and valid measure of performance. Toward this end, we need to evaluate the reliability of individual raters as well as the average rating. We will also need ways to estimate the number of raters needed to achieve a predefined reliability for any single performance. In practice an effective system should be able to identify inconsistent raters as well as performance that generate non-converging responses.

We begin by putting this reserch in the larger context of improving human and automated scoring in assessment. We argue that the classical theories of measurement are ill-suited for the new challenges. We

outline a theory of measurement that focuses on the scoring function that maps observable features to a confidence or likelihood function.

The second part of the paper attemps to identify a metric for measurement quality based on the theory. We start with measurement quality in the classical testing theory, focusing in particular on the intraclass correlation (ICC) family of measures (Bartko, 1966, 1976; Bonett, 2002; Fleiss & Cohen, 1973; Haggard, 1958; Lahey, Downey, & Saal, 1983; McGraw & Wong, 1996; Müller & Büttner, 1994; Shrout & Fleiss, 1979; Swiger, Harvey, Everson, & Gregory, 1964). We show that while such indices are useful in the part of our research that follow the traditional psychometric practices, they do not provide ready solutions to the problem we ultimately wish to solve.

The third part of the paper therefore outlines unique challenges in the definition and application of the notion of reliability in a rating study where human raters are dynamically allocated to scoring work products in order to maximize the quality and minimize the cost of rating. We propose an approach to evaluate the quality of rating and discuss its application in optimizing crowd-sourced human rating studies.

2 Toward a Theory of Scoring

Scoring or rating in the current context is the assignment of numbers (or symbols that can be numerically labeled) to an artifact. In this sense it is akin to the classical definition of measurement by S.S. Stevens, namely "the assignment of numbers to objects or events according to rules" (1951, p. 21). This definition makes no distinction between scoring by humans or by an algorithm – in fact one could argue that algorithmic scoring is a purer manifestation of the "rules" Stevens referred to.

2.1 Classical psychometric approaches

Subsequent theories of measurement highlight the functions (or rules in Stevens, 1951) that map from the subject to a number, although in trying to accommodate the measurement of abstract ideas they move toward the notion of latent variables. Carmines and Zeller (1979), for example, defines measurement as the process that links abstract concepts to empirical indicants, or "the particular sense data at hand" (Riley, 1963). Carmines and Zeller further noted that in order to link *observable response* (the "sense data at hand") on the one hand and the *underlying unobservable concept* on the other, there has to be a strong relationship between empirically grounded observations and the unobservable concepts. The warrant for this strong relationship – and hence the strength of the measurement model – comes from the *auxilary theory* that specifies the relationship between the concept and indicators.

This classical notion of measurement, however, does not suit problems in scoring complex artifacts by humans or by computer algorithms. Specifically,

- The notion of *observable responses* becomes murkey, because unlike in traditional tests where responses are explicitly and narrowly defined (e.g., marking option A), in the context of educational assessment artifacts to be evaluated often have to be characterized by a collection of (multidimensional and potentially infinitely large set of) features. A classic example is an essay, which can be decomposed into hundreds of linguistic features or more (**???**). What constitutes an *observable response* in this case becomes unclear. Instead, it is helpful to go back to Riley's (1963) notion of "sense data at hand." The features that describe the artifact must be *sensed* by the person or algorithm who does the rating. This is the basis for rating.
- While the idea of an *unobservable concept* may be useful for constructing a rubric, it has little utility in the context of rating artifacts once a rubric is defined. By any definition of measurement, once we are able to assign numbers or labels in a systematic and justifiable way, the *unobservable* is now measured or observed.

  - For those who wish to argue *measured* is different from *observed*, let's note that in order to claim an *observable response* is observed, you have to be able to systematically label it, which is logically equivalent to assigning a numerical value in the nominal scale, AKA measuring it.
  - This is in fact not a trivil point. In developing a scoring algorithm one routinely

- Hence in rating or scoring an artifact, the rubric takes primacy. A rubric is the "rules" in S.S. Stevens' definition of measurement. It also defines the sysmatic mapping from "sense data at hand" to the concept that used to be *unobservable* such that with the rubric, the concept becomes measured or observed.

  - Note that the rubric is **not** the *auxilary theory* that is required in classic theories of measurement in order to provide warrant to the measurement (Carmines & Zeller, 1979). A rubric is not a theory. Rather it is nothing more or less than a system of rules or a mapping from "sense data at hand" about an artifact to a numeric value that we shall call the rating or score for the artifact. Sometimes it is justified by theories; other times it may be putative.
  - A rubric can be called an operationalization of an auxilary theory, although this does little to clarify the relation. Operationalization is perhaps best understood as a set of putative rules justified on the basis of a theory but not necessarily uniquely predicted by it. In research operatinalization can sometimes be used as a hedge against situations where empirical (read: observed or measured) data do not turn out as predicted by the theory (suggesting that the measurement is ill-defined or of poor quality). In assessment a rubric is often revised when the mapping is found to be unsatisfactory. (**uh, what exactly am I arguing?**)

2.2 Functional approaches

A different take on the problem of measurement comes from the functionalism tradition known as the Brunswik's Lens theory (Brunswik, 1955, (**???**); Wolf, 2005).

2.3 A computational approach

Scoring or rating in educational assessment is similar to measurement in that both involve a systematic assignment of numbers (or logcial equivalents) to something of interest, be it an object, event, or abstract concept. It differs, however, from the classical notions of measurement in the following ways:

- The "thing" to be evaluated is a concrete but potentially complex artifact (e.g., an essay or a video recording of a performance), rather than some abstract latent concept. The assignment of numbers must be based in one way or another on the artifact, not on abstract notions. A scorer or rater of an educational artifact cannot justify her assignment based on "this is how I felt;" it must be supported by evidence linked to the artifact. A fair scoring process must be as transparent as possible; declairing the "thing" to be rated on as unobservable or latent does nothing to solve the problem at hand.

- Being tangible or perceptable doesn't imply that the artifact is *observable*, however. Rarely is an artifact in an educational assessment so simple that it can be quantified by a single number or classification (if so scoring is accomplished). We assume that an artifact can be distingished from other artifacts by a collection of descriptors, including, for example, *this essay uses the word 'sophistication'* or *the interviewee smiled at video frame 4000*. While Riley (1963) may call them "particular sense data at hand," we use the phrase *features* that is widely used in computational sciences to describe the potentially infinite number of ways to characterize an artifact. Features are **descriptive**, as opposed to **evaluative**, in that they are used to tell one artifact from other (potential) examples. Value assignment is done with a rubric (see below). By definition the features will not fit in a single dimension (or else a single feature or number is needed to distinguish the artifacts). Their internal structure can be complex (e.g., may require a complex graph to describe) and uncertain (e.g., dynamic), in which case the *i.i.d.* requirement in traditional linear statistical models will likely be violated.

- Scoring or rating in educational assessment is evaluative, although not necessarily in the sense of rank ordering. Rather than using the metaphor of measurement (implying unidimentionality and continuity), we see rating or scoring as foundamentally categorical, i.e., the score assigned is a label for a category. Assigning a score x to an artifact A is equivalent to the assessment statement, *Artifact* A *belongs to category* X, which can be evaluative if the category X implies value. We do not presuppose any structure or dimentionality of the categories (e.g., categories can overlap, or one can have one category for each integer between 1 and 100), so long as they jointly divid the space of all possibilities exhaustively. The collection of assessment statements can evaluate an artifact in a unidimensional scale, multidimentional profile, or any other ways language permits. (Footnote: hereafter we use scores, ratings, and assessment statements interchangeably because we see numeric scores as labels for a logical statement. There may be contexts in which we wish to distinguish them, e.g., in reporting.)
- The mapping from descriptive features to evaluative scores is defined by a rubric. The purpose of a rubric is to ensure the systemacity, consistency, and transparency in scoring. Ideally a rubric should be comprehensive and explicit – as operational as a computational algorithm. In practice, though, it is often written, intetionally or unintentionally, with much room for (human) interpretation. A rubric can be detailed (or explicit in the case of a computational algorithm) or vague, dimensional or holistic, rule-based or just showing prototypical examples, etc. We shall revisit the issue of the freedom for (human) interpretation of a rubric as it is a critical issue in developing an automated scoring algorithm. Let us conclude with what are not rubrics: an auxilary theory is not a rubric; a simple Likerd scale asking "how do you feel about A" is not a rubric; a procedure to sum up points to get a total score is not a rubric.

To summarize, scoring is the process of mapping descriptive features of an artifact to a numeric label of an evaluative assessment statement according to a rubric. With this in mind, let us define the scoring function, which is at the center of human rating and automated scoring.

2.4 Scoring function and scoring quality

Consider an artifact A (e.g., a multimodal recording of a performance to be evaluated) represented as a multidimensional feature vector $\tilde{\mathbf{A}}$ that is to be assigned a numeric value x_i accoring to a certain criterion ($r_i : \tilde{\mathbf{A}} \mapsto \Re$), that maps $\tilde{\mathbf{A}}$ to a real valued score by a rater i. We further assume that there exists a criterion r (aka a rubric) that defines the "true" mapping ($r : \tilde{\mathbf{A}} \mapsto \Re$), thus $x = r(\tilde{\mathbf{A}})$ is the "true score" of A according to r.

We can further speculate that the mapping occurs in two steps. The first is a mapping from the feature set to a probability distribution of possible scores, and a second step is taken to choose the value with the highest probability. Hence we can redefine the scoring function as:

$$r : \tilde{\mathbf{A}} \mapsto p(x)$$

where $p(x)$ is the likelihood of value x represented as a probability distribution. The rater always chooses x with the greatest likelihood.

The quality of x_i is intuitively the reduction of uncertainty about x when x_i is known. In other words, the posterior distribution of the true score should be narrowed after rater i rated.

$$p(x|x_i)$$

The above is construed differently from the classical testing theory due to the application we will pursue in Section 3.

- The artifact A and its feature set $\tilde{\mathbf{A}}$ are invariant. In automated scoring models the feature set will be extracted algorithmically and hence consistently. We also assume that all human raters perceive the features equally (e.g., all raters scoring a video performance can see the video and hear the speech), though they would weigh the features differently in scoring.
- Differences in ratings across raters are caused by differential mapping functions r_i. This is a departure from the general notion in the classical testing theory, where it is typically assumed that $x_i = x + e_i$ where e_i is $i.i.d.$. We push the individual differences from scores to the scoring function because an automated scoring function is essentially an algorithmic reliazation of such a scoring function. Furthermore, we anticipate idiosyncratic yet self-consistent scoring functions for individual raters in a crowd-sourced human scoring study. We need to model these rather than simply sweeping them under the rug of error variance.
- We construe the quality of a score in term of how it shed light on the "true" score, not necessarily on how numerically close it is to the true score. We do not assume the numeric value is on something more than a nominal scale. For example, x may represent a label for a statement. In this case, $x_i - x$ is undefined. However, $p(x|x_i)$ always is.

Our formulation has similarities with the functionalism tradition of Brunswik (Brunswik, 1955, (**???**)). Like Brunswik (1955) we see scoring – or learning to score – as a problem of multidimensional probabilistic learning task, where the rater strives to discover the optimal combination of observed proximal features of the (distal) artifact in order to achieve a high quality rating. On the other hand, we are less concerned with the ecological functionlism of Brunswik's theory than finding a framework to support a computational model. We also do not limit ourselves to the multiple regression models that have been popular in the literature (**???**, **???**).

3 Metrics of Quality of Scoring

3.1 Scoring quality in the classical testing theory

The issue of scoring quality has a long history in the classical literature on psychological testing. It is herefore a logical place to look for a solution. In addition, our immediate concern is that we have conducted a generalizbility study (**???**) where multiple raters rated all artifacts. We need to estimate the quality of the rating. Classical testing theory provides the convenient machinary.

We now review how quality is defined in classical testing theory frameworks. Two quantities are often used in classical testing theory to evaluate the quality of a measurement, namely reliability and validity, or in Muller and Buttner (1994), comformity and consistency, respectively. (Shold we discuss reliability in the generalizability framework, (Dimitrov, 2002)?)

We will come to focus on the ICC family indecies.

3.2 Forms of ICC

Numerous forms of ICCs have been identified in the literature (McGraw & Wong, 1996; Shrout & Fleiss, 1979). They can be derived from an ANOVA framework, depending on assumptions about the rating situation.

Let's start with a model presented by the classic McGraw and Wong (1996) paper. We shall focus on the derivation of the ICC(c, k) model.

More generally, the following classes of ICCs are often used in practice.

3.2.0.1 ICC(1,1)

A case where different raters rated each item, or an one-way ANOVA model.

3.2.0.2 ICC(2,1)

This correspond to a two-way ANOVA with random effects for both the rater and subjects. The rationale for designating raters as random effect, however, may be several (Hancock & Mueller, 2010, p. 151). One possibility is when the decisions of the study will be based on absolute scores (e.g., when there is an absolute cut-off score). Another common case is when the investigator wishes to use the fully-crossed G study results to estimate future D study where raters may not be fully-crossed; i.e., when we will in the future use a different rater to rate each subject.

3.2.0.3 ICC(3,1)

ICC(3,1) corresponds to a mixed-effect two-way ANOVA with random effect for subjects but fixed effects for the rater.

3.2.0.4 ICC with k raters

When the intended measure of scoring quality is based on the average of all k raters, we arrive at the following models.

3.3 Choosing a reliabilty measure

In the classic paper Shrout and Fleiss (1979) couched measurement selection in the ANOVA framework. In our case, where the primary interest is in the reliability of the mean score, we are in effect conducting what Shrout and Fleiss (1979, p. 426) called "a substantive study (D study)." The reliability of the mean rating is always greater than that of individual raters (Lord, Novick, & Birnbaum, 1968).

Shrout and Fleiss recommended that the number of observations (m) used to form the mean should be determined by a pilot reliability study (G study); see the next section. Alternatively this may be decided on "substantive grouns," meaning by practical considerations. Once a minimal number of individual raters is established, the reliability of the average rating of the m raters can be estimated using the Spearman-Brown forula and the ICC model for single rater reliability index, i.e., ICC(1,1), ICC(2,1) or ICC(3,1). **Huh? This doesn't make much sense**

Numerous authors have provided decision guidelines based on the Shrout-Fleiss model (e.g., Hancock & Mueller, 2010, p. 151). McGraw and Wong (1996) extended the Shrout and Fleiss family of ICCs and provided a decision tree for selecting the ICC measure for different use cases. According to the M-W framework (McGraw & Wong, 1996, p40), the appropriate model for our case, namely a two-way, random effect, average measure, consistency-based index, should be the ICC(c, k)

Muller and Buttner (1994) provided a decision tree to guide the selection of a reliability metric for a particular rating study. According to the decision tree, if we assume that raters are randomly selected, each rater rates each subject, and measurements are **not** exchanable, the correct measurements are Model B (k, P), Lin (2, P), or Kappa (2, NP). **Check to make sure they are ICC(2,k)**

In R ICC can be calculated using the function "icc" with the packages psy or irr, or via the function "ICC" in the package psych.

3.4 Determining the number of raters

The number of raters required to achieve a certain level of reliability is discucussed in (Shrout & Fleiss, 1979). Related, (Bonett, 2002) gave an approximation for deteriming the sample size requirements for estimating intraclass correlations with desired precision.

3.5 Limitations

Note, however, that these discussion assume that the number of raters will be fixed for different artifacts, at least missing values are none sysmatic.

In our application, however, we seek to assign different raters and different numbers of raters to each artifact. The assignment will be non-random. This violates the basic assumptions of the classical testing theory based derivations. Hence the results here are not applicable.

4 Optimizing Scoring Quality for Crowd-sourced Scoring

Our goal in this section is to design a scoring model where raters will be dynamically assigned to artifacts in order to maximize the quality of the scoring while minimizing a cost function for scoring. Results from the classical testing theory do not tell us how to dynamically allocate scoring resources to optimize scoring quality. Nevertheless they provide some basic parameters to start.

4.1 Defining a cost function

The cost of a rater i to score an artifact A_j is function of

- the time spent to score A_j
- the hourly rate of the rater i, assuming the rate is constant

though additional factor may come into play:

- the availability of i, i.e., the wait time for i to rate A_j
- the opportunity cost of i not rating other artifacts

But in general we assume the cost of the rater i is independent from the artifact A_j, or at least conditionally independent after some simple covariates are identified (e.g., the length of video recording).

4.2 Quality of rater

Let's attempt to define the quality of a rater. For a specific artifact A_j, the quality is the reduction of uncertainty about the true score distribution given x_i.

Over the population of A the quality can be defined over the joint distribution, which is strict. We could also define the quality based on the marginal distributions, which leaves open the rater-item interaction, where r_{ij} is different from r_i. We shall deal with the simple case first.

4.3 Optimizing for quality

Here is a simple strategy for optimization: Given a pool of m artifacts and n raters each having a scoring function $r_1 \dots r_n$ (or $r_{1j} \dots r_{nj}$ for the artifact A_j, where $j = 1 \dots m$), we set a minimal threshold of quality, and minimize the cost function.

References

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19*(1), 3–11. Retrieved from http://www.amsciepub.com/doi/abs/10.2466/pr0.1966.19.1.3

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, *83*(5), 762. Retrieved from http://psycnet.apa.org/journals/bul/83/5/762/

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, *21*(9), 1331–1335. Retrieved from http://www.researchgate.net/profile/Douglas_Bonett/publication/11266478_Sample_size_requirements_for_estimating_intraclass_correlations_with_desired_precision/links/5463d3d50cf2837efdb34670.pdf

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Vol. 17). Sage publications.

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement*, *62*(5), 783–801.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*. Retrieved from http://psycnet.apa.org/psycinfo/1974-04128-001

Haggard, E. A. (1958). Intraclass correlation and the analysis of variance. Retrieved from http://psycnet.apa.org/psycinfo/1959-02517-000

Hancock, G. R., & Mueller, R. O. (2010). *The reviewer's guide to quantitative methods in the social sciences*. Routledge.

Lahey, M. A., Downey, R. G., & Saal, F. E. (1983). Intraclass correlations: There's more there than meets the eye. *Psychological Bulletin*, *93*(3), 586. Retrieved from http://psycnet.apa.org/journals/bul/93/3/586/

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30. Retrieved from http://psycnet.apa.org/journals/met/1/1/30/

Müller, R., & Büttner, P. (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, *13*(23), 2465–2476. Retrieved from http://www.en.msc-epidemiologie.med.uni-muenchen.de/download/winter_term_2011_2012/epiresearchdesigns_articles/muellerbuttner02_12.pdf

Riley, M. W. (1963). Sociological research i. a case approach. Harcourt, Brace & World, Inc. New.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420. Retrieved from http://psycnet.apa.org/journals/bul/86/2/420/

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics.

Swiger, L. A., Harvey, W. R., Everson, D. O., & Gregory, K. E. (1964). The variance of intraclass correlation involving groups with one observation. *Biometrics*, 818–826. Retrieved from http://www.jstor.org/stable/2528131

Wolf, B. (2005). Brunswik's original lens model. *University of Landau, Germany*, *9*.