A RELIABILITY INDEX (a_i)

THAT ASSUMES HONEST CODERS AND VARIABLE RANDOMNESS

Xinshu Zhao

Hong Kong Baptist University

Paper presented at the 2012 annual conference

of Association for Education in Journalism and Mass Communication, Chicago.

Authors' Note

Xinshu Zhao is Chair Professor of Communication and a Co-Director of the Carter Center Initiative at Hong Kong Baptist University. He is also Cheung Kong Chair Professor of Journalism, Fudan University in Shanghai.

The author gratefully acknowledges the generous support of Prof. Rick Wong Wai-kwok, Vice President (Research and Development) of HKBU and Dr. Hui Huang, Chairman of Tenly Software and President of Shanghai Association of Information Sciences. The author also thanks Chi Yang and Guangchao Charles Feng for their skillful, efficient, and patient programming with PHP and R software, Augus Cheong, Zhaoyun Hu, Jianbin Jin, Shengqing Liao, Quan Liu, Fuyuan Shen, Qi Shen, Anbin Shi, Jun Tang, Qianfang Xia, and Weizhi Wendy Yin for organizing the data collection, and Jane Brown, Ke Deng, Xiaodan Fan, Zhongshi Guo, Sri Kalyanaraman, Jing Lucille Li, Jun S. Liu, Charles T. Salmon, Ning Mena Wang and Lixing Zhu for their comments. This study was also supported by HKBU Faculty Research Grant (2008 & 2009, Zhao PI), HKBU Strategic Development Fund (2009 & 2011, Zhao PI), and grants from Panmedia Institute (2010, Zhao PI) and ENICHD (R24 HD056670, Henderson PI).

Correspondence concerning this paper should be addressed to Xinshu Zhao at zhao@hkbu.edu.hk, Tel: 852-3411-7492, Fax: 852-3411-7375.

A RELIABILITY INDEX (a_i)

THAT ASSUMES HONEST CODERS AND VARIABLE RANDOMNESS

Abstract

The performances of six major indices of inter-coder reliability were evaluated against actual judgments of human coders in a *behavior-based Monte Carlo* (BMC) experiment. The correlations between the indices' estimated chance agreements (a_c) and the observed chance agreements (o_{ac}) turned out to be negative for Cohen's κ, Scott's π and Krippendorff's α, and mild although positive for Bennett et al's *S*, Perrault and Leigh's I_r and Gwet's AC_1. While each of the indices was designed to improve on percent agreement, each underperformed percent agreement (a_o) when estimating observed true agreement (a_t) in the BMC experiment.

The poor or negative correlations between the calculated *estimates* and the observed *estimands* question the validity of the *estimators*, namely the indices. The findings support the emerging theory that reliability indices available today assume dishonest coders who deliberately maximize chance coding, and they are therefore unsuitable for typical studies where coders perform chance coding involuntarily when the task is too difficult. A new index or indices are needed.

This manuscript also reports the effort to develop such a new index, *agreement index (a_i)*, which assumes honest coders and involuntary chance coding. Subsequent analysis shows

that a_i is void of the 23 known paradoxes that plague other indices. In the BMC experiment, the chance agreement estimated by a_i was by far the best predictor of the observed chance agreement between coders. Index a_i also outperformed percent agreement and all other six indices while predicting true agreements among the coders.

Empirical testing of theories and indices should continue, especially by different researchers using different methods, and so should the search for a better index. Until better evidences are available, however, researchers may refrain from using κ, π, and α, and add a_i as a reasonable measure of true agreements between two coders on a nominal scale. Online software has been provided at http://reliability.hkbu.edu.hk/ to facilitate calculation.

Key words: reliability, intercoder reliability, interrater reliability, agreement index, estimator, estimate, estimand, maximum randomness, variable randomness, behavioral Monte-Carlo experiment, BMC, simulation-augmented behavior experiment, SAB, kappa, alpha, pi.

A RELIABILITY INDEX (a_i)

THAT ASSUMES HONEST CODERS AND VARIABLE RANDOMNESS

Table of Contents

Author's Note 2
Abstract 3
Key words 4
Table of Contents 5
I. Available Indices and the Need for a New Index 8
II. Explicating Category, Distribution, and Difficulty 12
III. Testing Assumptions Using Behavior-Based Monte Carlo (BMC) Experiment 15
III.1. Rationale and Design of BMC Experiment 17
III.1.a. Behavior experiment is infeasible, while Monte Carlo simulation does not measure behavior 17
III.1.b. A behavior-based Monte Carlo (BMC) experiment 19
III.1.c. Identifying a task with a muddy gold standard 22
III.2. Execution of BMC Experiment 23
III.2.a. Manipulating category and difficulty at item level 23
III.2.b. Maintaining attention and minimizing interferences 24
III.2.c. Manipulating distribution at short-session level 25
III.2.d. Collecting and pairing coder responses 26
III.2.e. Simulating a 4X8X3 between-session experiment, based entirely on actual coder responses 27
III.2.f. Re-manipulating category, difficulty, and distribution at long-session level 29
III.2.g. Deciding number of subjects (sessions) per experimental cell 30
III.2.h. Measuring dependent and other independent variables at long-session level 30
III.2.i. Combining manipulation, behavior, and simulation 33
III.3. Existing Indices Performed Poorly, Because They Rely on Wrong Factors 34
III.3.a. Methodological check one: Short or long sessions did not affect distribution effect 34
III.3.b. Methodological check two: designed empty cells were not entirely empty 34

III.3.c. Methodological check three:
BMC experiment was orthogonal as designed 35
III.3.d. Distribution or category correlated strongly with estimated
chance agreements but not with observed chance agreement 36
III.3.e. Difficulty correlated positively with observed chance agreement
but not with estimated chance agreements 37
III.3.f. The estimate-estimand (a_c-O_{ac}) correlation
was negative for κ, π, or α and low for S, I_r, or AC_1 38
III.3.g. Major agreement indices did not improve on percent agreement 41
III.3.h. Summary of findings so far 43
IV. An Index, a_b, Under Black-White Randomness Assumption 44
V. Agreement Index, a_i, Under Mixed Randomness Assumption 51
VI. Evaluating a_i Against Paradoxical Scenarios and Experimental Data 64
VI.1. Agreement Index (a_i) Is Void of Known Paradoxes and Abnormalities 64
VI.2. Agreement Index (a_i) Performed Well in BMC experiment,
Because It Relies on Right Factors 65
VI.2.a. Observed disagreement is a reasonable indicator of difficulty 65
VI.2.b. Existing indices did not use observed disagreement
to estimate chance agreement 66
VI.2.c. Chance agreement estimated by a_i correlated highly
with observed disagreement and difficulty but
not with distribution or category per se 66
VI.2.d. Chance agreement estimated by a_i is the
best predictor of observed chance agreement 67
VI.2.e. Agreement index (a_i) is the best predictor of true agreement 67
VI.2.f. An on-line software for calculating a_i 67
VII. Conclusion 67
References 72
Figure 80
Tables 81

A RELIABILITY INDEX (a_i)

THAT ASSUMES HONEST CODERS AND VARIABLE RANDOMNESS

Indices of *intercoder reliability* have been often used to assess the quality of communication content studies. Researchers in other fields, such as psychology, education, sociology, and medical research, often use the same indices, where they were also referred to as *interrater reliability* or *agreement indices*.

Methodologists, however, disagreed over which index(es) of reliability are appropriate for general use. While Cohen's κ is by far the most popular across disciplines, many authors pointed out that it makes unrealistic assumptions about coder behavior, and therefore produces many paradoxes when used in typical research (Feinstein & Cicchetti; 1990; Grove, Andreasen, McDonald-Scott, Keller, & Shapiro, 1981; Lombard, Snyder-Duch, & Bracken, 2002; Zhao, 2011a; Zhao, Liu, & Deng, 2012b).

Scott's π (1955) is the second in popularity, and Krippendorff's α (1970a, 1980) has been regarded by communication researchers as the most sophisticated (Hayes & Krippendorff, 2007; Krippendorff, 2004b). But some argued that π and α make as many unrealistic assumptions and produce as many paradoxes as κ (Lombard et al., 2002; Zhao, 2011b; Zhao et al., 2012b). A Monte Carlo simulation found that the three indices behave almost identically to each other under many conditions (Feng, in press). A recent review of

22 indices found none of the existing indices satisfactory, and new indices based on more realistic assumptions were called for (Zhao, Deng, Feng, Zhu, & Chan, 2012a; Zhao et al., 2012b).

This manuscript begins with a brief review of the known limitations of the existing indices. I will then report an experiment using real human coders, which was also augmented by Monte Carlo simulation techniques, to empirically test these theories and criticisms. I will report that the major indices performed poorly, supporting the theories behind the criticism and the call for a new index that assumes more realistic coder behavior. I will report the efforts to develop such a new index, called *agreement index* (a_i). I will report that a_i is void of the known paradoxes, and it outperformed each major index in estimating the observed true chance agreement or the observed true agreement among human coders, who participated in the behavior-based Monte Carlo (BMC) experiment.

I. Available Indices and the Need for a New Index

Since Benini (1901), two types of intercoder reliability indices have been introduced. The first is the non-adjusted indices, including *percent agreement* (a_o, pre 1901), *Holsti's CR* (1969), *Osgood's* coefficient (1959) and Rogot and Goldberg's A_1 (1966). These indices assume that observed agreement contains no random chance coding, hence no need to adjust for it. As random chance coding is seen as given, many considered percent agreement "the

most primitive" (Cohen, 1960, p. 38), "inadequate" (Hughes & Garrett, 1990, p. 193), and "flawed" (Hayes & Krippendorff, 2007, P. 80). There has been a consensus among reliability experts that "percentage agreement should *not* be used … as an intercoder reliability estimation" (Hughes & Garrett, 1990, p. 187), leading to decades-long efforts to "account for" and "remove" chance agreement (Krippendorff, 1980, pp. 133-134; Riffe, Lacy, & Fico, 1998, pp. 129-130; Rust & Cooil, 1994, p. 2).

The second type is *chance-adjusted* indices. While the efforts to consider chance have been widely applauded, some authors argued or hinted that these indices' assumptions of coder behavior may be unrealistic (Grove et al, 1981; Lombard, Snyder-Duch, & Bracken, 2002; Riffe, Lacy & Fico, 1998, 2005; Rust & Cooil, 1994). These indices assume that coders deliberately maximize random chance coding, and limit honest coding to occasions dictated by chance, according to analyses by Zhao, Liu and Deng (Zhao, 2011a&b; Zhao et al., 2012b).

The chance-adjusted indices include three subgroups. The first is category-based indices. These indices assume that, as the number of categories increases, chance agreements decrease, and reliability increases, producing a classic paradox -- "empty categories increase reliability" (Scott, 1955; Zhao et al., 2012b).

Indices in the second subgroup estimate chance agreement as a function of distribution, which medical researchers refer to as "prevalence" (Feng, in press; Gwet, 2010;

Shrout, Spitzer, & Fleiss 1987; Spitznagel & Helzer, 1985). Some of these distribution-based indices are among the most popular, including Cohen's κ, Scott's π, and Krippendorff's α. While some consider this subgroup, especially α, the most sophisticated (Hayes & Krippendorff, 2007; Krippendorff, 2012), others argued that relying on distributions implies uncomfortable assumptions, e.g., coders apply pre-determined quotas (Zhao, 2011a,b; Zhao et al., 2012b). Consequently, π, κ and α produce more paradoxes and abnormalities than most of the other indices (Brennan & Prediger, 1981; Gwet, 2008, 2010, 2012; Zhao, 2011a&b; Zhao et al., 2012b).

The third subgroup has just one index, Gwet's AC_1, which is based on both category and distribution. The double base limits but not eliminates the negative impacts of distribution-related assumptions. The double base also brings back the troublesome impact of the category-related assumptions that the distribution-based indices successfully avoided (Zhao et al., 2012b).

A growing number of authors pointed out that chance agreement is affected by difficulty -- the higher difficulty leads to more chance coding, hene more chance agreement (Grove et al, 1981; Gwet, 2008, 2010, 2012; Riffe et al., 1998, 2005). Some argued that chance agreement is *not* a function of category (Scott, 1955; Zhao et al., 2012b), others argued it is *not* a function of distribution (Gwet, 2008, 2010, 2012; Zhao et al., 2012b). Some

called for new indices that use difficulty to estimate chance agreement (Gwet, 2008, 2010, 2012; Zhao et al., 2011a&b; Zhao et al., 2012b).

This study has three tasks. The first is to test some major assumptions implied in the published criticisms of the major indices and the call for a new index(es). A *behavior-based Monte Carlo (BMC)* experiment, which may also be called a *simulation-augmented behavior* (*SAB*) experiment, was designed for the test.

The second task is to develope a new reliability index based on three assumptions:

1) Chance coding often happens, and often produces chance agreements. This assumption is to avoid the main deficiency of the non-adjusted indices.

2) Coders do not deliberately maximize chance coding. Instead, they conduct chance coding involuntarily and often unknowingly. Consequently, chance agreement is not fixed at a certain maximum as most of the chance-adjusted indices assume. Some studies may produce no chance agreement, when the task is extremely simple and training exceptionally good, while others may produce maximum chance agreement, when the task is extremely difficult or the training exceptionally lacking. Most of the studies may fall somewhere between the two extremes. This assumption is to avoid a main deficiency of the chance-adjusted indices.

3) The amount of chance agreements, therefore, is a function of task difficulty, but not category or distribution per se. This assumption is to avoid another main deficiency of the chance-adjusted indices.

The third task is to test whether the new index achieves its design objectives, using again the data from the BMC experiment.

Obviously *category*, *distribution* and *difficulty,* are crucial concepts in our theorizing and analysis. The existing indices rely on category, distribution, or both, while we, among others, believe that difficulty is far more important. So let's explicate the three concepts before discussing empirical tests and mathematical derivations (c.f., Chaffee, 1991).

II. Explicating Category, Distribution, and Difficulty

Category is the number of choices available to a coder on a nominal scale. *Gender*, for example, typically has two categories, male and female, while *political party* may have two or more categories depending on country and time.

Distribution is the pattern of occurrences in each category, where "occurrences" are often expressed as percentages (Cohen, 1960; Feng, in press; Gwet, 2010, 2012; Perreault & Leigh, 1989). The concept has also been referred to as "frequency" (Gwet, 2008), "base rate" (Grove et al., 1981; Kraemer, 1979; Spitznagel & Helzer, 1985), or "prevalence" (Feng, in press; Gwet, 2010; Shrout et al., 1985).

In reliability literature, distribution is usually operationalized on a skewed-even continuum (Cohen, 1960; Gwet, 2008, 2010; Krippendorff, 1970, 1980; Scott, 1955). For example, on a binary scale, a 100% & 0% or 0% & 100% distribution is the most skewed, while a 50% & 50% distribution is the most even. So a distribution scale originally ranging from 0% & 100% to 100% & 0% needs to be "folded", making a 0% & 100% distribution equal to a 100% & 0% distribution, and a 10% & 90% distribution equal to a 90% & 0% distribution, etc. This study will follow this tradition.

Another tradition is to assume that coders' *reported distribution* is a good estimate of *target distribution* under coding (Cohen, 1960; Gwet, 2008, 2010, 2012; Krippendorff, 1970, 1980; Scott, 1955). It was based in part on this assumption that π, α, κ and AC_1 multiply the marginals of a contingency table to estimate chance agreement. The assumption has been widely accepted. Researchers conduct a research because they do not know the target distribution, and reported distribution is seen as the best indicator of the target distribution.

The target-report relation, however, may be more complicated, and it may depend on task difficulty and reported skew. When the task is easy, the coding tends to be accurate, and reported distribution tends to resemble the target distribution closely, whether the reported distribution is skewed or even. When the task is difficult, however, the coding tends to be random, and the reported distribution tends to be even, even when the target distribution is skewed. So a reported skewed distribution is more likely to resemble the target distribution

than a reported even distribution. Further, some argued that π, κ and α assume that reported distribution equals *marble distribution*, where "marble" refers to a physical, statistical, electronic, mental, or virtual device of probability that coders use to guide their chance coding (Zhao, 2011a&b; Zhao et al., 2012b). At least one author disagreed that such an assumption exists (Krippendorff, 2012). Our data will show that the assumption does appear to exist, and it is a major cause for the negative correlation between the chance agreement estimated by these indices and the chance agreement observed from our empirical data.

Distribution is more meaningful for a coding session with multiple items than a single item. By contrast, *category* and *difficulty* can be equally meaningful for a session or an individual item. This distinction has implications for our variable manipulation and data analysis, as I will explain later.

The concept *difficulty* may be defined broadly, as the joint consequence of all factors that cause the coding to be inaccurate. Thus defined, *difficulty* may include several dimensions:

1) *Task*: By their nature some tasks are more difficult than others. For example, deciding whether an advertisement contain Surgeon General's warning is easier than deciding whether a news story contains bias.
2) *Instrument*: Instrument refers to measures that researchers should take to enable the coders to accomplish a given task accurately. Examples of such measures

include sufficient incentive, organization, monitoring, appropriate categories, clear and appropriate questions, sufficient instruction, good training, adequate equipment, and good environment. Insufficient incentive or monitoring, inappropriate categories, unclear instructions, and poor training increases difficulty.

3) *Coders*: Some coders are less capable, less focused, or less motivated than others, which increase difficulty. Despite researchers' efforts, there are always variations among coders and within coders over time.

In the BMC experiment that I am to report below, I manipulated *task difficulty* so it ranges from very low to very high. By choosing a task that is almost self-explanatory, I fixed *instrument difficulty* at a very low level. But we still saw variations in *coder difficulty* between coders and over time.

III. Testing Assumptions Using Behavior-Based Monte Carlo (BMC) Experiment

Dozens of intercoder reliability indices have been introduced in over a century. Popping (1988) identified 39. Zhao and colleagues (2012a&b) reviewed 22 and quickly added a 23rd. Until recently, however, discussions of reliability indices relied on theoretical reasoning and mathematical derivation, and sometimes illustrated by a couple individual

examples. The last index introduced in this fashion was Potter & Levine-Donnerstein's redefined *Pi* (1999). Debates took place from time to, e.g., among Grove et al (1981), Kraemer (1979), Shrout et al (1987), Spitznagel & Helzer (1985), and Thompson & Walter (1988), between Krippendorff (2004b) and Lombard et al (2002), and between Krippendorff (2012) and Zhao et al (2012b). All sides relied on theoretical and mathematical analyses. We have not seen a systematic field study to build the empirical foundations for the indices and the debates.

Gwet (2008) took a large step forward. He used a Monte-Carlo simulation to support his mathematical analysis of four indices. 500 computer-generated samples, which I call "coding sessions," simulated a binary scale and two coders, although no coders were actually used. The simulated distribution within each session was fixed at 95% & 5%, and sample sizes were chosen at 20, 60, 80 and 100. Against a "'true' inter-rater reliability" defined by the author as the criteria, AC_1 showed smaller biases than Scott's π (1955), Cohen's κ (1960) or Guilford's *G* (Guilford, 1961; Holley & Guilford, 1964), while *G* is mathematically equivalent to Bennett, Alpert, and Goldstein's *S* (1954; cf., Zhao et al., 2012b). AC_1 also showed smaller variance than any of the other three indices. Feng (2012, in press) and Zhao et al (2012a) followed suit, by using larger Monte Carlo samples to explore various issues related to intercoder reliability.

Like any method, simulation has its limitations. While we need to test the effect of

category, difficulty, and distribution on coder behavior, pure simulation does not measure the actual behavior. It can simulate the behavior according to assumptions imposed by researchers, but the results cannot serve as tests of the assumptions.

What simulation cannot do, behavior experiment can. For example, experiment using human coders (subjects) can convincingly measure coder behavior. But a typical behavior experiment is infeasible in this case, which I will explain in more details below. So I combined experimental and simulation techniques to design a *behavior-based Monte Carlo* (BMC) experiment.

III.1. Rationale and Design of BMC Experiment

III.1.a. Behavior experiment is infeasible, while Monte Carlo simulation does not measure behavior.

Reliability studies require variables at the level of coding session. For example, percent agreement, distribution, Cohen's κ, Scott's π, or estimated chance agreement is meaningful only for a coding session with multiple target items coded by at least two coders. The unit of analysis must be coding session, not individual coder or single item.

Manipulating these variables at the level of coding session, however, are hardly feasible for a typical university researcher. We need a sufficient number of targets within each coding session (I chose N_t=100, where N_t is *number of targets coded in a coding session*)

and we need a sufficient number of coding sessions (I chose N_s=384, where N_s is the number of coding sessions involved in an analysis). The total number of data points would be much large than a typical controlled experiment or a content analysis study. The study I will report below, for example, manipulated four levels of category, eight levels of difficulty, and three levels of distribution, making it a 4X8X3 experiment with 96 cells. Following the conventional rule of 20 subjects (sessions) per cell, I would have needed 96X20=1,920 coding sessions with 100 targets and at least two coders for each session. At least 20 variables were involved in the analysis, which would mean 1,920 X 100 X 2 X 20 = 7,689,000 data points. A behavior experiment of this size would be too expensive and too difficult to implement with typically limited resources of university researchers.

More importantly, within each of the 96 cells, each independent variable would have a fixed value, e.g., two categories, extremely easy, and 99% & 1% distribution. Each coder would have to repeat the same task under the same condition 100 times per session. They are likely to get bored, tired, disinterested, or quickly figure out the answers to the subsequent questions. The idiosyncratic combination or interaction between a certain coder and a certain condition would have a large effect, leading to large measurement errors. Examples include a coder who is exceptionally good at the most difficult task was assigned the most difficult task 100 times, and a coder who gets easily distracted by irrelevant categories gets the highest number of categories 100 times.

III.1.b. A behavior-based Monte Carlo (BMC) experiment.

A two-stage experiment was designed To meet the challenge. The first was a *manipulation-behavior* stage. *Category* and *difficulty* were manipulated at the level of individual items, and *distribution* was manipulated at the level of 10-item short sessions. Human coders coded in short sessions, and each coder had a variety of category and difficulty levels which were randomly rotated within each session. While the short sessions and varying tasks helped to maintain interest and quality, the data from this stage were not directly usable for my purpose, because each session did not have one level of difficulty or category, but up to eight or four levels. Also each session had too small a sample (N_t=10) to be representative of typical coding session.

Hence the second stage, the *Monte Carlo* stage. I randomly sampled from the data collected at the first stage to simulate a 4X8X3 between-session experiment in which each session had 100 items (N_t=100), and was assigned a certain level of *category*, *difficulty*, and *distribution*.

Category is known to affect indices like Bennett et al's *S*, Perreault and Leigh's I_r and Gwet's AC_1 (Bennett et al, 1954; Feng, in press; Guttman, 1946; Scott, 1955; Gwet, 2008, 2010, 2012; Zhao et al., 2012b). Distribution is known to affect indices like Cohen's κ, Scott's π, Krippendorff's α, and Gwet's AC_1 (Brennan & Frediger, 1981; Feinstein & Cicchetti, 1990; Feng, in press; Perreault & Leigh, 1989; Rust & Cooil, 1994; Shrout et al.,

1987; Zhao, 2011a&b; Zhao et al., 2012b). Increasingly, however, reliability experts argue that difficulty is the main factor affecting chance agreement, therefore *should be* the main factor affecting a reliability index (Grove et al, 1981; Riffe et al., 1998 & 2005; Gwet, 2010, 2012; Zhao et al., 2012b).

So category, distribution and difficulty are the three manipulated variables. Other independent variables and dependent variables are measured from coders' responses, as I will explain below.

It is not a typical human behavior experiment. A typical human behavior experiment does not use Monte Carlo simulation (Montgomery, 2008). Yet this design shares two most valuable features of human behavior experiments, namely physical rather than simulated manipulation of independent variables and direct observation of human responses (Montgomery, 2008). In that sense, I call it a *simulation-augmented behavior* (SAB) experiment.

It is also not a typical Monte-Carlo (MC) experiment. A typical MC experiment is based on a population of random numbers defined by assumptions guided by relevant theories or practical needs (Liu, 2001). This SAB experiment was based on a population of observed human behavior in response to experimental manipulation guided by relevant theories and practical needs. Yet this design shares two most valuable features of a typical Monte Carlo experiment (Liu, 2001). One is the task, which is to solve a problem through

simulation when deterministic or other methods are infeasible. The other is a technique, which is repeated random sampling. So we may also call this design a *behavior-based Monte Carlo* (BMC) experiment.

The BMC design gave us the best of two methods, namely behavior experiments' ability to systematically record actual human behavior and Monte Carlo simulations' ability to efficiently generate large amount of high-quality data, thereby making the infeasible feasible.

Accordingly, I first experimentally manipulated *category* and *task difficulty* at the level of individual items and manipulated *distribution* at the level of small (10-item) sessions. I then measured coders' responses at the level of individual items. Using exclusively the data from the experimental manipulation and the coder responses, I constructed through simulation 384 coding sessions, each of which having 100 target items.

In the process, the three variables originally manipulated at the level of individual items or short sessions were re-manipulated at the level of long (100-item) sessions through Monte-Carlo-style repeated sampling. There were four levels of category (2, 4, 6, 8), eight levels of difficulty (1~8 pixels as the difference between two longest bars), and three levels of distribution (99%&1%, 75%&25%, and 50%&50%). As is shown in Table**1**, this is a 4X8X3 between-subject experiment, where "subject" is simulated coding session based on actual responses from human coders.

[Table 1 About Here]

Besides testing the effects of the three manipulated variables, I also wanted to assess the performance of the various reliability indices against systematically measured coder responses. Such assessment has not been seen in the literature. The ranges of the three independent variables were designed in part to set up a realistic context for the assessment.

III.1.c. Identifying a task with a muddy gold standard.

Reliability indices are standards against which empirical studies are evaluated. Now that the indices themselves are evaluated, we need ask coders a question that comes with a natural gold standard, namely an answer key that is agreeable by almost every reasonable person.

Most of empirical studies, however, have no answer keys -- had there been answer keys, no coding would have been needed. What researchers usually have is a varying degree of difficulty. When the task is extremely easy, the coders may all agree on an answer. When the task is extremely difficult, the coders may disagree as often as they agree. Most of the studies are somewhere in between, with some disagreements, but more agreements. To create a realistic environment and capture this key variable, we need varying and manipulable degrees of difficulty. That means we need to find a task question that has 1) a gold standard acceptable by the scientific community, 2) varying "muddiness" in the eyes of coders and, 3)

controllable "muddiness" in the hands of the researcher.

The task that I designed for my coders is to find the longest bar among several bars, whose lengths were manipulated by computer programming. This task meets the three requirements set above. The details are reported below.

III.2. Execution of BMC Experiment

My assistants programmed a website for the invited coders to participate online, at http://reliability.hkbu.edu.hk/. The site served as a platform that helped us to perform the following.

[Figure 1. About Here]

III.2.a. Manipulating category and difficulty at item level.

Each coder was asked to pick the longest bar from two, four, six or eight bars with various lengths (Figure 1). So the first manipulated variable, *category*, had four levels.

The second manipulating variable, *task difficulty*, was operationalized by changing the differences between the two longest bars, which are 200 pixels long, plus or minus a few pixels depending on the experimental manipulation. The smallest difference, representing the highest difficulty, is one pixel, which is the smallest controllable element on a computer screen. The largest difference, representing the lowest difficulty, is eight pixels. Hence

difficulty has eight levels. To facilitate analysis and interpretation, this variable was linearly transformed to a 0~1 scale where 1 represents the highest difficulty (1 pixel difference) and 0 represents the lowest difficulty (8 pixel difference), and other six differences proportionally spaced in between.

The distance between the two longest bars is fixed at 150 pixels to minimize the effect of distance. When there are four, six, or eight categories, all bars other than the two longest are fixed at 150 pixels long. So I will refer to the two longest bars "long bars" and the other bars "short bars." This design was to assure that the real competition was between the long bars, so that *category* effect does not confound or complicate *difficulty* effect. This design was also to test Scott's (1955) classic theory that Bennett et al's *S* increases with empty cells, which provided a main enticement for the development of Scott's π (1955), Cohen's κ (1960), and Krippendorff's α (1970, 1980).

III.2.b. Maintaining attention and minimizing interferences.

A number of measures were taken to maintain a high level of interest and attention by the coders. Each coding session was limited to 10 items. This was in part to eliminate clutter effect, which has been found to be an important factor affecting memory and attitude in various settings (Nan & Faber, 2004; Thorson & Zhao, 1997; Zhao, 1989, 1997). The number of category, difficulty level, location of the longest bar, and the location of the second longest bar (left or right of the longest bar) are randomly rotated within each session to minimize

effects of learning, tiring and boring, primacy and regency, other serial position, and interaction between coder idiosyncrasies and difficulty, category, position, etc. Prior studies have theorized or demonstrated that these factors affect memory and attitudes (Jeong, Tran, & Zhao, 2012; Li, 2010; Terry 2005; Thorson & Zhao, 1997; Zhao, Shen, & Blake, 1995).

III.2.c. Manipulating distribution at short-session level.

The random rotation described above also manipulated *distribution* within each 10-item session. The rotation often placed five longest bars at the left of the second longest bars, and another five longest bars at the right, producing a 5~5 distribution. The random chance also produced some 4~6, 3~7, 2~8 distributions, and a couple 1~9 and 0~10 distributions. As I operationalized distribution as skew, in later analysis the distribution scale is "folded," e.g., a 4~6 distribution is re-coded to equal a 6~4 distribution, and both were given a value 0.6. Similarly, a 5~5 distribution is represented by 0.5, and 0~10 and 10~0 distributions are both represented by 1.0.

As can be expected from a random process, far more cases concentrated in 0.5 and 0.6 distributions than in 0.9 or 1.0 distributions. After being folded, the distribution is skewed toward left (the 0.5 side). The skewed distribution violates an assumption of least square significance tests. It's even worse from the view of experimental design, as we need roughly equal number of cases in all conditions. Extremely skewed distribution is expected to affect several major indices significantly. The very small number of cases in those conditions

hampers our ability to test these effects. Further, κ, π and α are undefined with 1.0 distribution. A 0.99 (1~99 or 99~1) distribution would be good replacement, but unavailable in 10-item sessions.

These shortcomings are part of the reasons that our main analysis will not be on the original 10-item sessions, but instead on simulated 100-item sessions. During the simulation I will re-manipulate *distribution* to include 0.99, and I will be sure to have equal number of cases per distribution level. By contrast, the values of *category* (2, 4, 6 & 8) and *difficulty* (0~1) will remain unchanged during the simulation and re-manipulation.

III.2.d. Collecting and pairing coder responses.

At the end of each 10-item session, a coder may choose to stop or do another 10 items. The first 10 items coded by the first coder were preserved and presented to the next coder who logged on. The 10 items given to the two coders were identical in terms of category, difficulty, locations of the two longest bars, the serial order of the items, and all other factors within the control of the researcher. These 10 items coded by two coders constitute 10 pairs of coded items, and constitute 10 basic elements for our subsequently analysis. Once a 10-item list has been coded by two coders, it will not be coded again by a third coder as long as there is simultaneous logging on. When a coder chose to code 10 more items or when a new coder logged on, he or she may code 10 pre-preserved unpaired items or, if there are no unpaired items left, code 10 items newly generated by the computer. This process repeated

for all subsequent coders and 10-item sessions.

Occasionally more than two coders coded the same 10 items due to simultaneous logging on. In these cases I randomly selected two coders' responses for our analysis. There were also a couple cases one coder completed all 10 items while his or her partner completed less than 10 items. I excluded such incomplete information for the purpose of this analysis.

The data collection took place between March 22nd and May 9th, 2012. Undergraduate and graduate students, teachers, computer technicians, office workers, research assistants and other professionals from over 15 colleges and two research firms in America, mainland China, Hong Kong, Macau and Singapore participated in the coding as a part of their class exercise, research training, or work assignment. They registered 383 given names or web names. The total number of participating coders should be around 400, as some of the same names came from different cities and different organizations, hence likely represented different persons. The coders logged on 2,490 times from 53 cities in Asia, Europe, and North America. They coded a total of 22,290 items, of which 19,900 were successfully paired, producing 9,950 paired responses for our analysis.

III.2.e. Simulating a 4X8X3 between-session experiment, based entirely on actual coder responses.

I defined the 9,950 paired responses as the population for this study, from which I sampled repeatedly and randomly to construct a 4X8X3 experiment, through the following

procedure.

I first set the number of items within each session (N_t) to be 100. Then I drew my first sample with three conditions: category=2, difficulty=0 (easiest) and distribution=1&99%. As said earlier κ, π and α are undefined when distribution is 0&100% or 100&0%. From the 9,950 paired responses I randomly sampled without replacement 100 pairs that meet the category (K=2) and difficulty (d_f=0) requirements; to meet the distribution requirement of 1% & 99%, I sampled one item with the longest bar located at the left of the second longest bar, and 99 with the longest bar located at the right.

This constitutes my first simulated coding session. After the computer collected relevant information from the 100 pairs, I returned them to the population of 9,950. I changed one of the three factors, category, from two to four, while difficulty and distribution remain unchanged. Under the new conditions (K =4, d_f=0 and distribution=1&99%), I drew my second sample (coding session). I repeated the process for every combination of category and difficulty, which gave us a total of 4X8=32 samples (sessions). I then changed distribution to 25&75%, 50&50%, 75&25%, and 99&1%, and repeated the process, which gave us 32X5=160 samples (sessions). Because all inter-coder reliability are symmetrical with regard to distribution, I folded the distribution scale so that 1&99% and 99&1% are represented by the same score of skew (s_k) = 0.99, 25&75% and 75&25% are represented by s_k= 0.75, and 50&50% is represented by s_k= 0.5. Now *distribution* has three levels, extremely

skewed (s_k=0.99, 32X2=64 samples), moderately skewed (s_k=0.75, 32X2=64 samples), and even (s_k=0.5, 32 samples). So that every level has an equal number of samples, I repeated the process to draw another 32 samples for the even (s_k=0.5) distribution, which gave us a total of 160+32=192 samples.

As κ, π, and α can be heavily influenced by a small number of cases, especially when the distribution is skewed (Feinstein & Cicchetti, 1990; Hoehler, 2000; Lombard et al., 2002, Vach, 2005; Zhao et al., 2012b), and I had just 64 samples in the extremely skewed condition (s_k=.99), I was concerned whether the sample size was large enough. To stabilize the distribution effect while still give all conditions an equal weight, I repeated the entire process to double the number of samples for all conditions, which means the total number of samples is also doubled to become N_s=192*384.

III.2.f. Re-manipulating category, difficulty, and distribution at long-session level.

The structured and repeated sampling also re-manipulated category, difficulty and distribution. Its impact on the three variables differ somewhat due to the natures of each variable.

As category and difficulty had been manipulated at item level, their values remained unchanged after re-manipulation. For example, each item in an extremely easy and eight-category session had the same d_f=0 and K=8 as an individual item. They had scattered in many 10-item sessions and were gathered to this simulated 100-item session, with no change

to the values of difficulty or category.

Distribution was a bit different. It was a session-level variable and was originally manipulated at the level of 10-item session. When switching from 10-item session to 100-item session, for the reasons explained earlier, the values of distribution changed, from the original 0.5, 0.6, …&1.0 to 0.5, 0.75 & 0.99.

III.2.g. Deciding number of subjects (sessions) per experimental cell.

Using the terminology of psychological experiment, this is a 4X8X3 between-subject design with four subjects in each cell, where "subjects" are coding sessions. So one may ask why not follow the 20-subject-per-cell rule to sample more, to put 20 sessions into each cell. First, our main independent and dependent variables are all on numerical scales. The 4X8X3 characterization assumes categorical independent variables, and therefore may overestimate the number of samples needed. Second, the 20-subject-per-cell rule assumes individual subject as unit of analysis, while we have coding session, each of which aggregating 100 pairs of responses. Aggregated data tend to have smaller variance, which generates more statistical power and needs smaller N_s.

III.2.h. Measuring dependent and other independent variables at long-session level.

[Table 2About Here]

For the 100 paired responses sampled, I calculated the 22 variables listed in Table 2. Recall that, to minimize the effects of position and order, I randomized the locations of the

long bars relative to the short bars for every item (Jeong et al, 2012; Li, 2010; Nan & Faber, 2004; Terry 2005, Zhao, 1989, 1997; Zhao et al, 1995). To meet the design objective of having two main competing categories, during the calculation I treated each long bar as Category 1 or 2, regardless of its location on screen.

Observed agreement (a_o) and observed disagreement (d_o) were directly available in our data, which are also available in typical studies using two or more coders, raters, or diagnosticians.

Observed right agreement (a_r) is the number of cases for which both coders gave the correct answers divided by the total number of cases for each sample (N_t). Observed erroneous agreement (a_e) is the percent of cases that two coders agree but their answers are incorrect. The two measures are also discussed conceptually in Table 6. They should not be parameters of an index as they are usually unavailable in typical studies. This methodological study measured them to produce a couple "gold standards" against which the indices can be empirically evaluated.

The calculation of observed chance agreement (o_{ac}) needs some explanation. Chance agreement, by its nature, may be right or wrong. The erroneous agreement (a_e) is directly observed, which is a result of chance coding according to our starting assumption of no systematic error. The randomly right agreement is mixed with the systematically right agreement, therefore needs calculation.

Due to our design of two long bars and several (0~6) short bars, the chance agreement came from two types of random selection: between two long bars, and among all bars. When the later resulted in an agreement on the longest bar, we designate it as a_{ra}, right agreement from random choices among all bars. While a_{ra} can be calculated, [1] it will be clear soon that we will not need the calculation.

Some of the agreement on the second longest bar (a_{el}) also came from random selection among all bars. The amount is the same as those falling on the longest bar, which is a_{ra}. The rest (a_{el}-a_{ra}) came from random selection between two long bars. Because it is random selection between two, the same amount (a_{el}-a_{ra}) should fall on the longest bar. That means that right agreement from random choices between long bars is (a_{el}-a_{ra}).

Therefore, based on probability theory, observed chance agreement o_{ac} is calculated by taking the sum of the above:

$$o_{ac} = a_e + a_{ra} + (a_{el} - a_{ra}) = a_e + a_{el}$$

Observed true (non-chance) agreement (a_t) is the observed chance agreement (o_{ac})

[1] All agreements on the short bars are results of coders choosing randomly among all bars. Such agreement should spread evenly among all categories, and 1/K of which should fall on each bar, including the longest bar. If there are four categories (K=4), and agreement on the two short bars is a_{s4}, then (a_{s4}/2) is the amount of the randomly right agreement produced by coders choosing randomly among the four bars. Similarly, if there are six or eight categories, and a_{s6} and a_{s8} represent the agreement on the four or six short bars, then (a_{s4}/4) and (a_{s4}/6) are the randomly right agreement resulted from random selection among all six or eight bars. So the total amount of right agreement resulted from random selection among all bars is a_{ra}=(a_{s4}/2)+(a_{s4}/4)+(a_{s4}/6).

minus the observed agreement (a_o):

$$\boldsymbol{a_t = a_o - o_{ac}} \qquad (1)$$

The four variables, o_{ac}, a_t, a_e, and a_r constitute a set of "gold standards" against which various indices can be assessed. None of the four is usually available in typical studies.

III.2.i. Combining manipulation, behavior, and simulation.

In the long-session data, which are the main basis of our subsequent analysis, the sessions and coders were both partially simulated. For example, in any 100-item session, the randomly selected responses could have come from up to 200 different coders. It was not a pure simulation because it's based mainly on actual behavior of real human coders, but not on purely theoretical assumption. It was not purely real because no such 100-item sessions actually took place.

I mentioned earlier that random rotation of categories, difficulties, positions etc. minimized the effect of individual coders and other idiosyncrasies. The partial simulation through random selection and reassembling further minimized the impact of individual coders or other idiosyncrasies attached to the original coding. It therefore made the data more stable and more representative of "typical" sessions, coders, and responses.

In a typical behavior experiment, independent variables are manipulated, and human responses are observed from real human participants in real experimental sessions. In a typical Monte Carlo experiment, however, all manipulation, participants, responses, and

experimental sessions are simulated. This BMC experiment has a mixture of manipulation, participation, and simulation. Some independent variables were physically manipulated at item level, then re-manipulated through simulation at session level. Other independent variables and all dependent variables were gathered from the actual responses of real human coders participating in real coding sessions. But the coders and sessions, as said, were reassembled through partial simulation.

III.3. Existing Indices Performed Poorly, Because They Rely on Wrong Factors

III.3.a. Methodological check one: Short or long sessions did not affect distribution effect.

I mentioned that, while the values of *category* and *difficulty* remained unchanged during the re-manipulation, values of *distribution* were changed. An underlying assumption was that distribution is not expected to affect chance coding whether a coding session is short or long, observed or simulated. I first checked this assumption at the level of short sessions (N_s=984). Distribution (s_k) was not correlated with observed chance agreement (o_{ac}, r=0.015, t=0.470, p=0.638) or percent agreement (a_o, r=0.004, t=0.135, p=0.893). Comparing them with the counterpart correlations (r=-0.023, p>.05, and r=-0.044, p>.05) at the level of long sessions shown in Table 3, I conclude that my assumptions are consistent with the data.

III.3.b. Methodological check two: designed empty cells were not entirely empty.

Recall that I made the long bars clearly longer than the short bars, so that the real competition would be between the two long bars. I therefore hoped for largely empty cells, especially empty "agreement" cells, for the short bars. They did not turn out to be as empty as I hoped. On average 2.86% of choices fell on the short bars, which was broken down to 1.11%, 1.93% and 5.53% for respectively four, six, and eight categories. As expected, agreement on short bars is much lower, at an average of 0.45%, and broken down to 0.04%, 0.12%, and 1.18% for respectively four, six, and eight categories. Although not exactly zero, these numbers are still small and therefore did not show a clear effect on our subsequent data analysis, theory testing or index development. [2]

III.3.c. Methodological check three: BMC experiment was orthogonal as designed.

The upper left corner (Columns A~C, Rows 1~3) of Table 3 shows, the correlations between the three manipulated variables. These zero correlations verify our orthogonal design.

Having checked on these methodological issues, I now report our main findings at the level of long sessions, which are summarized in Table 3. Note that the cells involving a_i is about a new index I will introduce in the next section. In this section I focus on all other cells that are about coder behavior and the six existing indices.

[Table 3 About Here]

[2] Interestingly, while category appears to have a small positive effect on choices and agreement on the short bars, it has a somewhat negative effect on choices and agreements on the second longest bars. Probably because the second mechanism is a bit stronger, the overall correlation between category and chance agreement is weakly negative. Overall, our data show that the effect of category on chance agreement is weak.

III.3.d. Distribution or category correlated strongly with estimated chance agreements but not with observed chance agreement.

The six coefficients at the upper right corner (Columns D~F; Rows 1&2) tested two assumptions implied in the criticisms of the existing indices, which are that distribution (skew) or category per se does not affect coder judgment (Grove et al, 1981; Scott, 1955; Gwet , 2008, 2010; Zhao, 2011a&b; Zhao et al., 2012b). Five of the six correlation coefficients were near zero and statistically non-significant, supporting the assumptions.

The only statistically significant correlation among the six was between *category* (*K*) and *observed chance agreement* (o_{ac}, r = -.138**). Further analysis revealed that the negative correlation was due to a drop in agreement on the second longest bar when there were eight categories (25.36% vs 28.94% as the average for two to six categories).[3] As agreement on short bars increased with category (as I reported above), and agreement on the second longest bar also increased between two and six categories (26.45%, 29.95%, 30.41% for respectively two, four, and six categories), the -.138 correlation does not appear to indicate a consistent effect of category per se on chance agreement.

Although actual chance agreement was not affected by distribution or category per se, major indices' estimated chance agreements were affected by one or the other (Feng, 2012, in

[3] Although the negative sign coincides with the prediction of three category-based indices, S, I_r and AC_1, which also predicts a negative K-a_c correlation, this -.138 correlation does not lend any support for the three indices, for two reasons. First, the observed process that produced the -.138 correlation is not predicted by any theories behind the three indices. Second, this observed K-o_{ac} correlation (r = -.138**) is too weak to justify the much stronger K-a_c correlations (r = -.813***~-.929) theorized by the three indices. In other words, the correlations bear the same sign by empirical coincidence, but not by theoretical necessity.

press; Grove et al, 1981; Perreault & Leigh, 1989; Rust & Cooil, 1994; Scott, 1955; Gwet, 2008, 2010; Zhao, 2011a&b). *S* and I_r depend on category, κ, π and α depend on distribution, and AC_1 depends on both category and distribution, according to a mathematical analysis of the formulas (Zhao et al., 2012b). Group II (Rows 5~10) of Columns A & B of Table 3 shows support for these observations and analyses. Chance agreements estimated by *S* and I_r correlated strongly and negatively with category (*r*=-.929, A6, A7). Chance agreements estimated by π, κ and α correlated positively and substantially with skew (*r*=.659~.661, B8~B10). Chance agreement estimated by AC_1 correlated quite strongly with category (*r*=-.813, A5) and relatively weakly with skew (*r*=-.197, B5). The last two correlation coefficients suggest that Zhao et al's (2012b) characterization of AC_1 may need a revision: while AC_1 is double-based, it depends on category far more than distribution.

III.3.e. Difficulty correlated positively with observed chance agreement but not with estimated chance agreements.

While each index's estimation correlated with skew or category when it should not, it did not correlate with difficulty in the way it should. The argument that task difficulty increases the chance coding was supported by the positive and relatively strong correlation between difficulty and observed chance agreement (*r*=.765***, E3). One may note that difficulty was also a good predictor of percent disagreement (r=.882***, D3), true agreement (*r*=-.880***, F3), and percent agreement (*r*=-.882***, C12), demonstrating again its

importance.

The estimations by the major indices, however, correlated minimally or even negatively with difficulty. Group II of Column C (C5~C10) shows that the correlations between difficulty and expected chance agreement was zero for the category-based indices (*S* and I_r, r=.000, C6 & C7), close to zero for the double-based index (AC_1, r=.095, C5), and negative for the distribution-based indices (π, κ, and α, $r \leq$-.351***, C8~C10). These findings support Gwet's (2008, 2010) criticism that the indices' before him had not taken difficulty into account. Nevertheless, the minimal correlation between difficulty and AC_1's estimation (r=.095, C5) may also disappoint Gwet (2008, 2010), whose AC_1 was designed to take difficulty into account.

III.3.f. The estimate-estimand (a_c-O_{ac}) *correlation was negative for κ, π, or α and low for S,* I_r, *or* AC_1.

As some might expect by now, these indices' estimations of the chance agreements showed up as poor *estimates* of the observed chance agreement (Group II, Column E). The highest correlation is a mild r=.273*** for AC_1, followed by *S* and I_r (r=.146***). The correlations for π, κ, and α were even negative (r=-.388*** and r=-.390***).

A statistical procedure is an *estimator* that produces *estimates* to approximate its *estimand*, which is the target phenomenon under estimation (Lehmann & Casella, 1998). A perfect estimator produces a positive and perfect estimate-estimand correlation. A good

estimator produces a positive and strong correlation. A negative estimate-estimand correlation is unacceptable. A negative and statistically significant correlation is alarming, as it suggests that the estimator routinely reports the opposite of the reality.

As mentioned earlier, the estimated chance agreement (a_c) is the most important element that defines each index. These indices remove the estimated chance agreements from the observed agreement in order to produce an estimated reliability. The negative correlation means that κ, π, and α regularly remove a large amount of "chance agreement" when the actual amount is small, and regularly remove a small amount when the actual amount is large.

The negative correlation is due to a mismatch between coder behavior assumed by the indices and the coder behavior observed in this experiment, especially in relation to reported distribution. A reported skewed distribution is assumed to indicate that coders have drawn from an equally skewed distribution of marbles. As a more skewed marble distribution produces more matches of marble color, and coders are assumed to code randomly when the colors match, a more skewed reported distribution is assumed to indicate *more* random coding (Perreault & Leigh, 1989; Rust & Cooil, 1994; Zhao, 2011a&b; Zhao et al., 2012b).

Typical coders, such as the coders in this study, do not code this way. They code honestly. When one codes randomly, it's not because the marble colors match, but because the task, the situation, or his or her condition is too difficult. Random coding produces more

even, not more skewed, results, according to elementary probability theory. So a more skewed reported distribution indicates *less* random coding (Rust & Cooil, 1994), which contradicts the three indices' estimates, hence the negative correlations.

This also explains the negative correlation between difficulty and the three indices' estimated random agreements (Cells C8~C10 in Table 3). In our data more difficult task produces *more* random coding, hence more even reported distribution, which the three indices see as an evidence of *less* random coding under the maximum randomness, predetermined quota, and trinity distribution assumptions (Zhao, 2011a&b; Zhao et al., 2012b).

This finding may appear even more alarming if we consider that in about half a century the three indices have used as the most authoritative indices of intercoder reliability across disciplines. They have been used as screeners in various stages of research process, from topic identification, protocol development, measurement selection, to publication. The negative estimate-estimand correlation and the underlying mechanism imply that the three indices –

1) favor studies, protocols, measures, and instruments that report more even distribution, even when the reports have been generated largely at random, and

2) disfavor studies, protocols, measures and procedures that report more skewed distribution, even when the reports have accurately reflected the underlying target distribution.

The application of the three indices on tens of thousands of published, unpublished, and unfinished studies might have portrayed a world that look more even than it actually is. In communication research, for example, there may have been overestimates of rare phenomena and underestimates of common phenomena, due to the widespread use of π and α. More worrisome would be in medical research, where prevalence of rare diseases may have been inflated, while the prevalence of common conditions may have been deflated, due to the enduring use of κ.

Reported distribution, which is not shown in Table 3, is important for understanding the three indices. This variable and its impact on the three indices deserve a more thorough analysis in a separate study.

III.3.g. Major agreement indices did not improve on percent agreement.

Column F of Group II lists the correlations between *observed true agreement* (a_t) and various indices. The highest are r=.849*** and r=.831*** for AC_1 and S, and the lowest is r=.559*** for π, κ and α. The reasonably high correlations do not necessarily indicate AC_1 and S as good indices of intercoder reliability. Although each chance adjusted index was meant to be an estimator while the true agreement (a_t) is the estimand, theses particular estimate-estimand correlations are not necessarily the sharpest differentiators between indices, for a couple reasons –

All major indices use the same Equation 7 to remove chance agreements, a_c. The

indices differ from each other in how to estimate a_c. In Equation 7, a_c is subtracted in the nominator and again in the denominator. The two subtractions have the opposite effects; the first reduces an index and the second increases it by a usually smaller amount. The partial offsetting reduces impact of a_c, making percent agreement (a_o) the dominant factor in the equation and on the indices. As all indices estimate a_o exactly the same and a_o is highly and positively correlated with observed true agreement a_t (r=.917***, F12, Table 3), all indices appear positively and at least moderately correlated with a_t. But that was not due to the unique features of any index, namely estimated chance agreement a_c, but due to percent agreement a_o, which each chance-adjusted index was designed to improve on.

Therefore, although we want a higher correlation between an index and observed true agreement a_t (Column F, Group III), all these correlations are inflated by a_o that dominates every index, hence not the most effective differentiator between indices. The correlation between an index's expected chance agreement a_c and the observed chance agreement o_{ac} (Column E, Group II) is a sharper differentiator.

So if an index-a_t correlation (Column F, Group III) looks barely acceptable while the counterpart a_c-o_{ac} correlation looks completely unacceptable (Column E of Group II), the index, such as π, κ, or α, may be completely unacceptable. If the former looks reasonably good while the latter looks marginally acceptable, the index, such as AC_1 and S, may be marginally acceptable.

One way to properly interpret an inflated indicator is to compare it with the inflator as a benchmark. As percent agreement a_o is the inflator, the a_o-a_t correlation (r=.917***, F12) becomes a proper benchmark. But there is more important reason for a_o as the benchmark, that is, each of the six indices adjusted for chance for the stated purpose of improving on a_o.

A comparison of the index-a_t correlations (r=.559~.849) with the benchmark a_o-a_t correlation (r=.917) shows no improvement. While one could argue that the benchmark is so high that significant improvement would be difficult, there is no justification for a significant impairment, such as the reduction in correlation from r=.917 to r=.559. Unfortunately the largest impairment came from the most popular indices, namely κ, π, and α.

III.3.h. Summary of findings so far.

Our assumption checking verified the following:

1) The design features of our experiment met our objectives and expectations in general.
2) Distribution or category per se did not affect observed chance agreement, which implies that we should not rely on either of the two to estimate chance agreement.
3) Each of the major indices relied on distribution, category per se, or both to estimate chance agreement.
4) More difficult tasks produced significantly more chance agreement in our data,

which implies that an index's estimation of the chance agreement should also be positively correlated with difficulty. More difficult task should lead to a higher estimation of the chance agreement.

5) Each of the major indices available now had a negative or near zero correlation with difficulty.

6) As a result, the major indices' estimated chance agreements were not positively and highly correlated with the observed chance agreement.

7) The above findings support the call for a new index that uses difficulty, but not distribution or category per se, as the main factor affecting chance agreement.

IV. An Index, a_b, Under Black-White Randomness Assumption

This section reports our first attempt to develop an index based on more realistic assumptions. A more comprehensive typology of coder agreements and disagreements was built as conceptual and theoretical foundation. A good typology is crucial for this type of work, as it 1) selects appropriate dimension(s) for classifying types (Zhao, 2002a; 2004a&b, 2007a); 2) provides an inclusive list to include all types (Zhao, 2002b, 2007b); and 3) sets mutually exclusive division(s) between types (Zhao, 2002b, 2007a). If a typology fails on any one of the three tasks, the resulted theory, calculation, or formula is likely to err.

Based on the first typology I built a new index. But this "new" index turned out to be a mathematical equivalent of an existing index. Tracing back the steps, I identified a cause: the underlying typology was not comprehensive enough. After the typology was revised and expanded, a truly new index was developed, which I will discuss in the next section.

I started with the assumption that all cases are either difficult, which leads to chance coding, or easy, which leads to systematic coding. Gwet (2008, 2010, 2012) made the same assumption. This is a black-white version of the variable randomness assumption. A more complicated version will be discussed in the next section.

Coding can be divided into systematic or random (Krippendorff, 1970b, 2008). Honest, diligent, and consequently accurate coding is by definition systematic. Such behavior produces desired systematic agreements. Cheating, incorrect instruction and equipment failure are among those that produce undesired systematic coding, agreements, and disagreements.

Based on this analysis, Table 4 depicts a typology of two dimensions 1) systematic or chance coding, which produces 2) agreements or disagreements. Systematic coding is further divided into desired and undesired.

[Table 4 About Here]

Researchers should do their best to eliminate undesired systematic coding. Data should be void of significant influences from such coding when they are analyzed. If such

influences are not eliminated, they are hard to estimate statistically, if possible at all. So I will follow all other indices to assume no undesired systematic miscoding, which means no systematic disagreements (D_v in Table 4) and no systematically erroneous agreements (A_v) (Krippendorff, 1970b, 2008; Zhao et al., 2012b).

Systematic and accurate coding should always produce intercoder agreements (A_a), but never disagreements (D_a). When there is a disagreement, both coders cannot be right, and the instrument cannot be accurate. So D_a by definition should be zero.

$$\boldsymbol{A_e = D_e = D_a = 0} \qquad (\ 2\ )$$

Now that two of the three types of disagreements are zero, the remaining type, chance disagreements (D_c), should constitute all disagreements (D_o):

$$\boldsymbol{D_c = D_o} \qquad (\ 3\ )$$

Chance coding (C) produces chance agreement (A_c) and chance disagreement (D_c), which according to probability theory tend to be equal to each other when the number of cases is sufficiently large:

$$\boldsymbol{C = A_c + D_c} \qquad (\ 4\ )$$

$$\boldsymbol{A_c = D_c = D_o} \qquad (\ 5\ )$$

It means that chance agreement (a_c) and chance disagreement (d_c) are expected to be equal to each other, and both equal the observed disagreement:

$$\boldsymbol{a_c = d_c = d_o} \qquad (\ 6\ )$$

Here $a_c \equiv A_c/N$ and $d_c \equiv D_c/N$, where N is the total number of cases analyzed.

After obtaining a_c, almost all popular indices, such as Bennett, Alpert, & Goldstein's *S,* Scott's π, Cohen's κ, and Krippendorff's α, use Eq. 7 to estimate the agreement index. Note that a_o is the observed agreement, a_c is the estimated chance agreement, and r_i is an reliability index:

$$\boldsymbol{r_i = \frac{a_o - a_c}{1 - a_c}} \qquad (7)$$

Recent analyses showed that the formula assume *maximum randomness,* that is, coders are assumed to always maximize chance coding, and conduct honest coding only when marble colors turn out in a certain pattern, e.g. mismatch. As the assumption is inconsistent with typical coder behavior, some argued that the indices should rarely be used if ever (Zhao, 2011a&b; Zhao et al., 2012b).

The numerator in Equation 7 is problematic, as a_c is estimated under the maximum randomness assumption. The denominator is also problematic. An agreement index is a percentage figure, for which the denominator defines the reference scale. The reference scale can affect an index as much as the numerator, a_o-a_c. For example, if the reference scale is 1, an a_o-a_c=0.4 produces an index value 0.4; but if the reference scale is shrunk to 0.5, the same a_o-a_c=0.4 produces an index value 0.8. By subtracting a_c from 1, Equation 7 shrinks the reference scale by an amount a_c, which is the chance agreement based on maximum randomness assumption. That means a maximum amount of chance coding is assumed to

have taken place before honest coding, and that is a major defect of the chance adjusted indices in use today (Zhao et al., 2012b).

Following this reasoning, I will revise Equation 7. I will still subtract a_c in the nominator, but will not do so in the denominator. Hence an agreement index a_b for binary scales based on a black-white variable assumption:

$$\boldsymbol{a_b = a_o - a_c = a_o - d_o} \qquad (\ 8\ )$$

Or we may estimate the true agreement A_b by removing chance agreement (A_c) from observed agreement (A_o):

$$\boldsymbol{A_b = A_o - A_c = A_o - D_c = A_o - D_o} \qquad (\ 9\ )$$

Under the variable random assumption, Eq. 8 does not shrink the reference scale, which means it divides by one, that is, does not divide.

Let's extend this to multiple categories with two coders. Let K be number of categories and C be the number of cases chance coded:

$$\boldsymbol{A_c = C\frac{1}{K * K} * K = \frac{C}{K}} \qquad (\ 10\ )$$

As $C=A_c+D_c$ (Equation 4) and $D_c=D_o$ (Equation 5), Equation 10 becomes:

$$\boldsymbol{A_c = \frac{A_c + D_c}{K} = \frac{A_c + D_o}{K}} \qquad (\ 11\ )$$

Solving Equation 11 for A_c, we have the following for multiple categories:

$$\boldsymbol{A_c = \frac{D_o}{K - 1}} \qquad (\ 12\ )$$

and

$$\boldsymbol{a_c = \frac{d_o}{K-1}} \quad (13)$$

Given that a_b=a_o-a_c (Equation 8), we have:

$$\boldsymbol{a_b = a_o - \frac{d_o}{K-1}} \quad (14)$$

To illustrate the assumptions behind reliability indices, Zhao et al. (2012b) provided a scenario for each index they reviewed. Here I provide a scenario for a_b to lay bare its assumptions, which I will call Black-White Scenario:

1. Coders place *K* sets of marbles into an urn, where *K* equals the number of coding categories. Each set has an equal number of marbles and has its own color. The coders agree on which color represents which category. Following Zhao et al. (2012b), I use "marble" to refer to any physical, virtual or mental element of equal probability, and "urn" to refer to any real or conceptual collection of the elements.
2. They take a target to be coded. Here *target* is anything under coding, such as an advertisement, a news story, a patient, etc.
3. Together the two coders decide whether the target is easy or difficult to code, and they always reach an agreement. If easy, they will code the target as it is, and go back to Step 2 to code another case. If difficult, they will go to Step 4.

4. One coder draws a marble randomly from the urn, notes the marble's color, and puts it back. He will code the target following the marble color according to the pre-determined color-category scheme.
5. The other coder does the same.
6. The coders repeat Step 2 and the subsequent steps, and end the coding session when they have thus "coded" all targets.

Comparing this *Black-White Scenario* with *Bennett Scenario* (Zhao et al., 2012b), we see that the assumptions behind a_b are very different from those behind Bennett et al's *S.* Nevertheless, as d_o=1-a_o, Equation 14 can be re-written as:

$$\boldsymbol{a_b = a_o - \frac{1 - a_o}{K - 1} = \frac{a_o K - 1}{K - 1}} \qquad (15)$$

Note that Bennett et al (1954) sets a_c as a function of K:

$$\boldsymbol{a_c = \frac{1}{K}} \qquad (16)$$

When a_c is inserted into Equation 7, we have Bennett et al's *S:*

$$\boldsymbol{S = \frac{a_o - \frac{1}{K}}{1 - \frac{1}{K}} = \frac{a_o K - 1}{K - 1}} \qquad (17)$$

Comparing Equation 17 with Equation 15, we can see a_b is mathematically the same as *S*, even though they are based on very different assumptions. Different assumptions led to the same formula, suggesting the following:

First, different assumptions led to the same index. Under the maximum-randomness assumption, Equation 7 overestimates chance agreement, and subtracts the inflated amount from the numerator, hence suppresses the index. But the same assumption requires us to subtract the same amount from the denominator, which inflates the index. The two effects offset each other to produce an index identical to that of under the black-white variable-randomness assumption.

Second, the black-white assumption may not be a sufficient improvement over the assumptions behind Bennett et al's *S*. The assumption is closer to actual coding situations, therefore an improvement in conceptualization, but not enough to also produce an improvement in actual computation. We need a computation under even more realistic assumption, which I will call mixed-random assumption.

V. Agreement Index, a_i, Under Mixed Randomness Assumption

Eqs. 8, 13, and 15 derive an agreement index based on a *black-white* assumption. A case is either sufficiently simple, so both coders code it completely systematically and accurately, or it is sufficiently difficult, so both coders code it completely randomly. The assumption has two statistical implications:

1) The two coders' chance coding has the same pattern. So their chance judgments that cause disagreements are the mirror images of each other. It would mean that each pair of off-diagonal cells equal each other, e.g., $J_{12}=J_{21}$, $J_{13}= J_{31,}$ and $J_{23}=J_{32}$ in Table 5.

2) Consequently, each coder's chance coding is evenly distributed across categories. That means chance disagreements (D_c), which equals observed disagreements (D_o), are also evenly distributed. If we tabulate the coders' judgments as shown in Table 5, it would mean $D_{11}=D_{12}= D_{13,}$ $D_{21}=D_{22}= D_{23}$, etc.

[Table 5 About Here]

Together, the two assumptions expect all off-diagonal cells to be evenly distributed, that is, to have equal number of cases in such cells. Deviations from this pattern are assumed to be random variations, hence routinely ignored by Equations 13 & 15.

Alternatively, we may assume *mixed randomness*. Besides the sufficiently simple and extremely difficult, many cases are somewhere in between. A case may be sufficiently simple for one coder, who codes it systematically, yet too difficult for another, who codes it randomly. Or one coder codes with sufficient care and attention, while the other does so while being board or tired. Or one coder may pay attention some times, while get tired other times. Or he may pay half attention, or find the task somewhat difficult, so he codes partially randomly and partially systematically. Or, when there are three or more categories, a coder

may find the differences between some categories clear while others unclear, so he differentiates accurately between some categories while randomly between others.

[Table 6 About Here]

Table 6 presents a typology of coding based on mixed-randomness assumption. The typology modifies the black-white typology in Table 4 by inserting a new Column 3 for mixed coding, including mixed agreements (A_m) and mixed disagreements (D_m). Earlier we set three criteria for a good typology. 1) It selects appropriate dimension(s) for classifying types (Zhao, 2002a; 2004a&b, 2007a); 2) It provides an inclusive list to include all types (Zhao, 2002b, 2007b); and 3) It sets mutually exclusive division(s) between types (Zhao, 2002b, 2007a). If a typology fails on any one of the three tasks, the resulted theory, calculation, or formula is likely to err. If we compare Tables 4 and 6, we may say that the black-white typology failed to provide an inclusive list.

Under the mixed randomness assumption, chance coding is still seen as a main factor behind all disagreements, but no longer the only factor. Systematic coding also contributes to disagreements. So disagreeing patterns of two coders are not necessarily the mirror images of each other, and disagreeing judgments of each coder is not necessarily evenly distributed across categories. The uneven pattern of disagreements is not disregarded as random variation. Instead, they are seen as useful information for estimating chance agreement.

Consider two scenarios: 1) Both coders code completely randomly; and 2) One coder codes completely randomly, while the other systematically and accurately. With a binary scale, both scenarios are expected to generate a 50% disagreement (d_o) and a 50% agreement (a_o). But the agreements in Scenario 1 are expected to be half right and half wrong, while the agreements in Scenario 2 should be all right and no wrong.

Chance agreement usually has been seen as a unitary concept. Rarely if ever have scholars discussed different types of chance agreements. The mixed randomness assumption challenges this view. Chance functions differently in mixed coding (Column 3 of Table 6) and the purely chance coding (Column 4 of Table 6), therefore produces different types of chance agreements.

[Table 7About Here]

A target may be coded by both coders accurately, or by one coder accurately while by the other randomly, or by both randomly. When both code randomly, they may randomize between all categories or between some categories, e.g., one coder's difficulty is between three categories while the other's is between two categories. For a three-category scale, each coder has four possible ways of coding, and two coders have 16 possible combinations in Table 7.

When both coders code randomly, called pure chance coding, it produces either *erroneous chance agreement* (a_e), e.g., two coders agree that a target belongs to Category 1

when it actually belongs to Category 2, or *non-consequential chance agreement* (a_n), which is a type of *correct chance agreements* (a_r), e.g., two coders agree that a target belongs to Category 1 when it indeed belongs to Category 1. Mixed coding can produce only one type of chance agreement, *mixed chance agreement* (a_m), which is also a type of *correct chance agreements* (a_r). Mixed coding means that at least one coder codes accurately. So it cannot produce erroneous agreements.

[Table 8 About Here]

The unitary view of chance coding needs to be replaced by a multi-element view, and the concept of *chance agreement* should be replaced by the concept of *chance-affected agreements*. Table 8 shows a typology of chance-affected agreements based on this theory.

Some may argue that all chance-affected agreements must be removed. Others may focus only on the erroneous chance agreement, and emphasize that the other two types, mixed and non-consequential agreements, produce correct results, albeit by accident. A midway approach is to remove all pure chance agreements, including non-consequential and erroneous ones. Together, the three types may constitute an "acceptable range" between what must be removed (erroneous agreements) and what could be removed (all chance-affected agreements).

Ideally, we would like to estimate each of the three types, so that researchers may pick and choose what to remove. But typical reliability data contain only observed

agreement, observed disagreement, distribution pattern, and scale category, and not much more. So we may not be able to precisely calculate each of the three types. To be realistic, we may have to set a target range between erroneous agreements and all chance-affected agreements, and work out a "fuzzy" index whose will hit somewhere in the range.

Since we are estimating a type of agreement, that is, chance agreement, our instinct might direct our attention to the observed agreements (a_o). But a_o may not tell us much. The "agreement row" in Table 6 shows why -- its composition is too complex. Three cells are not empty – Honest and accurate agreement (a_a), mixed agreement (a_m), and, finally, purely chance agreement (a_c). We want to estimate a_c. But it is mixed with the other two types of agreements, especially a_a, which comes from systematic coding. Systematic behavior does not follow the expected probability of a random process, hence cannot be estimated using statistical means. The composition of a_o varies from study to study. Under Grove-Riffe Scenario, if the task is extremely easy, $a_a$could be 100%, so a_o=a_a=1, and a_m=a_c=0; if the task is extremely difficult, it could be 0%, so a_o=a_c, and a_a=a_m=0. We probably have no way of estimating the directly which part of $a_o$occupies how large a share. While we always want more accurate agreements (a_a), it is also the main un-estimable part.

Parallel to observed agreement (a_o) is observed disagreement (d_o), which may provide more useful information for estimating chance agreement. In Table 6, directly under a_a is d_a, disagreement resulted from both coders coding accurately, which is a logical impossibility. If

both coders code accurately, they have to agree with each other, so disagreement is impossible, which means d_a is zero by definition

Further, deliberate and systematic disagreement (d_e) is assumed to have been prevented or eliminated by the time of data analysis. Therefore the observed disagreement (d_o) have only two components left, mixed disagreement (d_m) and pure chance disagreement (d_c). So $d_o=d_m+d_c$, which implies that composition of observed disagreement (d_o) is simpler than observed agreement (a_o), making d_o more useful for estimating chance agreement.

As both d_m and d_c are chance affected, the distribution pattern of the observed disagreement (d_o) may be a good indicator of the distribution pattern of chance coding. For example, in Table 9, of all the disagreements, Coder 1 placed 51.9% in Category 2 (d_{12}=.296), while Coder 2 placed 37.0% in the same category (d_{22}=.370). Based on this, we may estimate that, of all the chance-affected coding done by Coder 1, 51.9% resulted in his choosing Category 2. Similarly, of all the chance-affected Coding done by Coder 2, 37.0% resulted in his choosing Category 3. The product of the two, .519*.37.0, estimates the percent of chance-affected coding that resulted in the two coders agreeing on Category 2 by chance.

[Table 9 About Here]

In general, we use D_{1c}/D_o to estimate the proportion of the chance-affected coding that Coder 1 puts into Category C, and D_{2c}/D_o to estimate the proportion of the chance-affected coding that Coder 2 puts into the same category. Further, we use the product

$D1c*D2c\ /D_o^2$ to estimate the probability that the two coders' chance-affected coding producing chance-affected agreements on Category C, which I denote as c_{cc}:

$$\boldsymbol{c_{cc} = \frac{D_{1c}}{D_o} * \frac{D_{2c}}{D_o}} \qquad (\ \mathbf{18}\ )$$

For example, Eq. 19 calculates c_{c1}, the probability that Coder 1 and Coder 2 agree on Category 1 from their chance-affected coding.

$$\boldsymbol{c_{c1} = \frac{D_{11}}{D_o} * \frac{D_{21}}{D_o}} \qquad (\ \mathbf{19}\ )$$

Adding c_{cc} probability across all categories, we get c_c, the probability of two coders producing chance-affected agreements from their chance-affected coding:

$for\ D_o > 0$

$$\boldsymbol{c_c = \frac{D_{11}}{D_o} * \frac{D_{21}}{D_o} + \frac{D_{12}}{D_o} * \frac{D_{22}}{D_o} + \cdots + \frac{D_{1c}}{D_o} * \frac{D_{2c}}{D_o} = \sum_{c=1}^{\infty} \frac{D_{1c} * D_{2c}}{D_o D_o} = \frac{\sum(D_{1c} D_{2c})}{D_o D_o}} \qquad (\ \mathbf{20}\ )$$

D_o=0 indicates zero disagreement, hence no evidence of chance-affected coding, therefore no chance-affected agreement, hence c_c should be defined as zero,

$$\boldsymbol{c_c = 0 \quad if\ D_o = 0} \qquad (\ 21\ )$$

Note a potential discrepancy between the stated objective and the actual calculation reflected in Equation 20. I set out to estimate chance-affected agreement, which includes mixed agreement that is not only affected by chance coding but also by systematic and accurate coding. Equations 19 and 21, however, assume the process is purely by chance. As systematic coding does not follow any probability rule, there is no easy way of estimating it. Consequently c_c may not precisely estimate all chance-affected agreements. Instead, it

measures only a portion of it. As a result, we may underestimate chance-affected agreements when we use c_c as a main factor in our estimation.

When disagreements are evenly distributed between two categories and do not appear in any other categories, c_c reaches its maximum at 1/2. This can happen when there are no other categories or when other categories contain no disagreements. In Table 9, it means all disagreements are in one pair of off-diagonal cells, and are evenly distributed between the two cells, e.g. $J_{21}=J_{12}$ and $J_{1c}=J_{2c}=0$ for all other off diagonal cells.

From 1/2, c_c decreases in two ways: (1) When distribution between the two cells of the pair becomes uneven; e.g. $J_{21}\neq J_{12}$. It reaches its minimum, zero, when all disagreements are in one cell. (2) When disagreements also appear in other categories, that is, when $J_{1c}>0$ or $J_{1c}>0$ for some other off diagonal cells; c_c approaches zero when disagreements are evenly distributed in a large number of categories. Hence,

$$\mathbf{0 \leq c_c \leq ½} \qquad (\ 22\ )$$

For a binary scale with two coders, c_c would be:

$$\boldsymbol{c_c = \frac{D_{11} * D_{21}}{D_o * D_o} + \frac{D_{12} * D_{22}}{D_o * D_o} \quad if\ D_o > 0} \qquad (\ 23\ )$$

$$\boldsymbol{c_c = 0 \quad if \quad D_o = 0} \qquad (\ 24\ )$$

Focusing on conceptual meaning, c_c, the probability of two coders producing purely chance agreements from their chance-affected coding, may also be defined in terms of A_f, the number of chance-affected agreements, and D_f, the number of chance-affected disagreements:

$$\boldsymbol{c_c = \frac{A_f}{A_f + D_f}} \quad (24)$$

Under the assumption that all observed disagreements, D_o, is chance affected, D_o=D_f, Eq. 24 becomes:

$$\boldsymbol{c_c = \frac{A_f}{A_f + D_o}} \quad (25)$$

Rearranging Eq. 25, we have:

$$\boldsymbol{A_f = D_o \frac{c_c}{1 - c_c}} \quad (26)$$

Defining a_c=A_f/N, and dividing both sides of Equation 26 by N, we have

$$\boldsymbol{a_c = d_o \frac{c_c}{1 - c_c}} \quad (27)$$

Deriving from Equation **8** (a_b=a_o-a_c) and Equation 27, we have an agreement index, a_i, based on mixed-randomness assumption:

$$\boldsymbol{a_i = a_o - a_c = a_o - d_o \frac{c_c}{1 - c_c}} \quad (\mathbf{28})$$

Because $0 \leq c_c \leq 1/2$ (Inequality 22).

$0 \leq c_c/(1- c_c) \leq 1$, hence $0 \leq d_o \leq d_o * c_c/(1- c_c)$

With this in mind, we compare Eqs. 27 and 28 with Eqs. 6 and 8. We can see that a_c estimated under the black-white assumption is equal to or larger than a_c estimated under the mixed randomness assumption; hence a_b under the black-white assumption is equal to or smaller than a_i under the mixed randomness assumption.

Under black-white randomness assumption, Equations 13 & 15 treat d_o as the only factor, hence treat the two scenarios as if they are the same. Under mixed randomness assumption, the size of d_o is insufficient to estimate chance agreement. We need to also consider the distribution pattern of d_o.

Under the mixed randomness assumption, the size of d_o is still the most important factor affecting a_c and a_i,[4] but the distribution pattern of disagreements becomes another factor. When the pattern is completely even between two categories, a_c=d_o, and a_i is simply the difference between a_o and d_o. When the disagreement distributes to more categories, or distributes unevenly, a_c is smaller than d_o, and can be as small as zero; and a_i is larger than the difference between a_o and d_o, and can be as large as a_o.

Replacing c_c in Equation 28 with the right side of Eq. 20, we have:

$$\boldsymbol{a_i = a_o - d_o * \frac{\sum(D_{1c}D_{2c})}{d_o * d_o - \sum(D_{1c}D_{2c})} \quad (for\ d_o > 0)} \qquad (\ \mathbf{29}\ )$$

$$\boldsymbol{a_i = 1\ (for\ d_o = 0)}$$

Like all other agreement indices, a_i assumes a certain pattern of coder behavior (See Zhao et al., 2012b), which is described in the following a_i Scenario:

1. The coders take a target to be coded. Here *target* is anything under coding, such as an advertisement, a news story, a patient, etc.
2. Each coder decides whether the target is easy or difficult to code. If he finds it easy, he codes the target as it is, then goes back to Step 1 to code another case. If

[4] This is equivalent to saying "a_o is the most important factor," since a_o=1- d_o.

he finds it difficult, he puts it aside to be coded later, and then go back to Step 1 to code another case. The two coders repeat Steps 1 and 2 until all targets have been coded by both coders, put aside by both coders, or coded by one and put aside by another, at which point they go to Step 3.

3. For each difficult target, the coder also judges whether it is difficult between two categories, three categories, etc., and between which of the categories. For each category, he calculates the percentage frequency of cases he judged difficult, which we call "difficulty distribution."
4. Each coder prepares a physical or mental urn. He fills the urn with physical or mental marbles using "difficulty distribution" as the distribution for marble colors. Each coder draw marbles from his urn with replacement to code every target that he judged as difficult.

In comparison with the scenarios assumed by other major indices (see Zhao et al., 2012b), a_i Scenario appears closer to typical coder behavior, hence should be a more reasonable estimate of true agreement from typical studies. Nevertheless, a_i still contains a portion in Steps 3-5 that seems to deviate from typical coder behavior. A revised Steps 3-5 would be closer to typical coder behavior:

3. For each target that a coder decided to be difficult, he judges whether it is difficult between two categories, three categories, etc., and between which of the categories.

4. Each coder prepares *K*-1 urns, where *K* equals the number of coding categories. The first urn has *K* sets of marbles, where each set has an equal number of marbles and has its own color. The second urn has *K*-1 sets of marbles, and so on, and the last urn has two sets of marbles. The coders agree on which color represents which category. Here, again, "marble" refers to any physical, virtual or mental element of equal probability, and "urn" refers to any real or conceptual collection of the elements.

5. Each coder draws marbles from an appropriate urn with replacement to code every target that he judged as difficult. In the "appropriate urn" the marble colors equal the categories between which the coder judges as difficult. For example, if a coder judges the decision to be difficult between three categories, he would draw from an urn with three marble colors.

Index a_i is not based on this revised scenario because I have not found a way to model the behavior pattern.

VI. Evaluating a_i Against Paradoxical Scenarios and Experimental Data

VI.1. Agreement Index (a_i) Is Void of Known Paradoxes and Abnormalities

Researchers have reported unexpected behavior of various intercoder reliability indices (Brennan & Prediger, 1981; Cohen, 1960; Feinstein & Cicchetti, 1990; Grove et al, 1981; Gwet, 2008, 2010; Hayes & Krippendorff, 2007; Kraemer, 1979; Krippendorff, 2004b; Lombard et al, 2002; Riffe et al., 2005; Scott, 1955; Spitznagel & Helzer, 1985; Zwick, 1988). Zhao et al. (2012a&b) reviewed 23 indices and identified 23 paradoxes and abnormalities

We applied a_i to the situations or examples related to each of the 23 paradoxes and abnormalities, and performed calculations where needed. During this scrutiny a_i did not display any sign of the paradoxical or abnormal behaviors of the other indices that Zhao and coauthors (2012a&b) described. With regard to avoiding the known paradoxes and abnormalities, a_i is a clear improvement over each of the 23 indices.

VI.2. Agreement Index (a_i) Performed Well in BMC experiment, Because It Relies on Right Factors

Since Benini (1901), dozens of reliability indices have been introduced without the support of large-scale empirical data, simulated or observed. Gwet (2008) started a new practice by supporting his AC_1 with a Monte-Carlo simulation using computer generated data. This study took a step further. I conducted a controlled experiment that manipulated category and difficulty, and measured responses. I then sampled from the individual responses to

simulate an experiment with hundreds of coding sessions and various distributions.

The data support the prior criticism of major indices that they rely on the "wrong" factors, namely category or distribution, and do not rely on the "right" factor, namely difficulty. Consequently, these indices' estimated chance agreements do not predict the actual chance agreement accurately, and the most "sophisticated" indices, namely κ, π, and α, predicts the opposite of the observations.

In this section, I extract and analyze the data from the same experiment to evaluate a_i and compare it with the other indices. The results are inserted into Table 3 for easy comparison.

VI.2.a. Observed disagreement is a reasonable indicator of difficulty.

While difficulty appears to be the most important factor affecting chance agreement, it is not directly observable in typical studies. To produce a new index based on difficulty, we needed a surrogate measure. Through theoretical analysis, I identified observed disagreement, which is readily available in typical studies with two or more coders, diagnosticians, or other raters. Cell D3 of Table 3 shows that observed disagreement (d_o) is a good surrogate, as it is positively and highly correlated with difficulty (r=.882***).

VI.2.b. Existing indices did not use observed disagreement to estimate chance agreement.

The six major indices do not appear to take advantage of the observed disagreement.

As shown in Cells D5~D10, none of the indices' estimated chance agreement is highly and positively correlated with disagreement. The highest correlation is a statistically non-significant r=.088 for Gwet's estimation. The correlations for κ, π and α are negative and substantial, r=-.472*** and r=-.475***, showing again a source of these indices' inaccuracies – not relying on what they should rely on.

VI.2.c. *Chance agreement estimated by a_i correlated highly with observed disagreement and difficulty but not with distribution or category per se*.

If the major indices do what they shouldn't and don't do what they should, a_i appears to do just the opposite, as is shown in Table 3. It is not correlated with distribution (r=-.008) and is mildly although negatively correlated with category (r=-.186***). This negative correlation was due to the same factor that produced the negative correlation between category and observed chance agreement (r=-.138**) discussed earlier. The closeness of the the two correlations with each other indicates that the effect of category on a_i is about right in direction and in magnitude.

Further, agreement index (a_i) successfully used observed disagreement (d_o) as a surrogate measure of difficulty (d_f) -- a_i's estimated chance agreement (ai_{ac}) has an r=.946*** with d_o and an r=.853*** with d_f (D11 and C11 of Table 3). Among the estimations of all indices, ai_{ac} is by far the best predictor of d_o and d_f. The second best is AC_1 estimation which has an r=.088 with d_o and r=.095 with d_f.

VI.2.d. Chance agreement estimated by a_i is the best predictor of observed chance agreement.

It should not be surprising by now that ai_{ac}, the chance agreement estimated by a_i, is by far the best predictor of the observed chance agreement o_{ac}. The correlation between ai_{ac} and o_{ac} is r=.745***, in comparison with the second highest, r=.273*** between Gwet's AC_{ac} and o_{ac}.

VI.2.e. Agreement index (a_i) is the best predictor of true agreement.

It should also not be surprising that a_i is the best predictor of the observed true agreement, a_t, producing an r=.921***, which is slightly higher than the correlation between percent agreement (a_o) and a_t, which has an r=.917***. Among all the correlations in this block (Group III of Column F), these are also the only two that are above 0.9.

VI.2.f. An on-line software for calculating a_i.

To facilitate the calculation of a_i, we have developed an online software that is now available http://reliability.hkbu.edu.hk/. It was initially programmed by Guangchao Charles Feng and Chi Yang, then further developed and now maintained by Tenly Software.

VII. Conclusion

The performances of six major indices of intercoder reliability were evaluated against actual judgments of human coders in a behavioral Monte Carlo (BMC) experiment. It's

discovered that true random agreement is a function of difficulty but not distribution or category per se.The major indices' estimations, however, rely heavily on distribution, category per se, or both, but not on difficulty. Consequently, the correlations between the indices' estimated chance agreements (a_c) and the observed chance agreement (o_{ac}) were mild or negative. When estimating true agreement (a_t), all six indices underperformed percent agreement (a_o), which the indices had been designed to improve on.

The poor or negative correlations between the calculated estimates and the observed estimands question the validity of the estimators, namely the indices. The findings support the emerging theory that reliability indices available today assume dishonest coders who deliberately maximize chance coding, and they are therefore unsuitable for typical studies where coders perform chance coding involuntarily when the task is too difficult. We should suspend the use of some indices, especially κ, π and α, until new evidences emerge to show when they can play a positive role. A new index or indices are needed.

This manuscript reports the effort to develop such a new index, agreement index (a_i), which assumes honest coders and involuntary chance coding. Unlike all other indices, the chance agreement estimated by a_i is not a function of distribution or category per se, but a function of observed disagreement. The BMC experiment has shown that observed disagreement is an excellent surrogate indicator of task difficulty.

Subsequent analysis shows that a_i is void of the 23 known paradoxes that plague other indices. In the BMC experiment, the chance agreement estimated by a_i was by far the best predictor of the observed chance agreement between coders. Index a_i also outperformed percent agreement and all other six indices while predicting true agreements among the coders.

No index is perfect. None will be. Empirical testing of theories and indices and the search for a better index will continue. They should continue, especially by different researchers using different methods. Different results may emerge with different tasks, designs, coders, instructions, etc. Until new evidences or better indices are available, however, a_i should be considered a reasonable measure of true agreement between two coders on a nominal scale. To facilitate calculation, we have developed an online software available at http://reliability.hkbu.edu.hk/.

It's hoped that researchers and methodologists will apply a_i to actual data and compare it with other indices, especially Cohen's κ (1960), Scott's π (1955), Krippendorff's α (1970a, 1980), Bennett et al's *S* (1954), Perreault and Leigh's I_r (1989) and Gwet's AC_1 (2008, 2010, 2012). It was after years of usage and scrutiny that deficiencies of κ and some other indices became known. It may also take time and application before we know the nature of a_i.

The experimental design may be another contribution of this study. In a typical Monte Carlo experiment, the universe of data is defined by theory-guided assumptions assembled to simulate a certain conditions of the real world. The computer-calculated results may simulate human responses to the simulated conditions, but may not test the theories and assumptions used to set up the simulation.

In this behavior-based Monte Carlo (BMC) experiment, the universe of data is defined by actual human behavior in response to the experimental conditions physically manipulated by the researcher. The results may test the theories and assumptions not used to set up the experiment.

What makes this BMC experiment different from typical human behavior experiment is that BMC data were not directly analyzed, but instead reorganized through Monte Carlo sampling before analysis, so that the data are more representative of real environment under which human coders actually behave. In that sense, this BMC experiment is more a behavior experiment than a Monte Carlo experiment. The design may also be useful in other studies where behavior needs to be observed at individual level but analyzed at group level, such as organizational studies.

References

Benini, R. (1901). *Principii di Demongraphia: Manuali Barbera Di Scienze Giuridiche Sociali e Politiche* (No. 29). Firenze, Italy: G. Barbera.

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communication through limited response questioning. *Public Opinion Quarterly, 18,* 303-308.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement, 41, 687-699.

Chaffee, S. H. (1991). *Explication*. Sage Publication.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Feinstein, A. R., & Cicchetti, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43,* 543–549.

Feng, G. C. (in press). Factors Affecting Intercoder Reliability: A Monte Carlo Experiment. *Quality and Quantity*.

Feng, G. C. (2012). Indexing versus Modeling Intercoder Reliability. Doctoral Dissertation, Hong Kong Baptist University.

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry, 38,* 408–413.

Guilford, J. P. (1961, November). *Preparation of item scores for correlation between individuals in a Q factor analysis.* Paper presented at the annual convention of the Society of Multivariate Experimental Psychologists.

Guttman, L. (1946). The test-retest reliability of qualitative data. *Psychometrika, 11,* 81-95.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61,* 29–48.

Gwet, K. L. (2010). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters (2nd ed.). Gaithersburg, MD: Advanced Analytics, LLC.

Gwet, K. L. (2012). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters (3rd.). Gaithersburg, MD: Advanced Analytics, LLC.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1,* 77–89.

Holsti, O. R. (1969). *Content Analysis for the Social Sciences and Humanities.* Reading, MA: Addison-Wesley.

Holley, W., & Guilford, J. P. (1964). A note on the *G*-index of agreement. *Educational and Psychological Measurement, 24*, 749-753.

Hughes, M. A., & Garrett, D. E. (1990). “Intercoder Reliability Estimation Approaches in

Marketing: A Generalizability Theory Framework for Quantitative Data." *Journal of Marketing Research, 27*(2), 185-195.

Jeong, Y., Tran, H., & Zhao, X (2012). How Much Is Too Much? The Collective Impact of Repetition and Position in Multi-Segment Sports Broadcast. *Journal of Advertising Research, 52*(1), 87-101.

Kraemer, H. C. (1979). Ramifications of a population model for kappa as a coefficient of reliability. *Psychometrika*, *44,* 461–472.

Krippendorff, K. (1970a). Bivariate agreement coefficients for reliability of data. *Sociological Methodology, 2,* 139-150.

Krippendorff, K. (1970b). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30,* 61-70.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology.* Newbury Park, CA: Sage Publications.

Krippendorff, K. (2004a). *Content Analysis: An Introduction to its Methodology* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Krippendorff, K. (2004b). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30, 411–433.

Krippendorff, K. (2007). Computing Krippendorff's Alpha Reliability. *University of Pennsylvania Scholarly Commons,* http://repository.upenn.edu/asc_papers/43

Krippendorff, K. (2008). Systematic and random disagreement and the reliability of nominal data. *Communication Methods and Measures, 2(4),* 323-338.

Krippendorff, K. (in press, 2012). A dissenting view on so-called paradoxes of reliability coefficients. *Communication Yearbook 36*.

Lehmann, E. L., & Casella, G. (1998). Theory of Point Estimation. NY: Springer-Verlag New York.

Li, C. (2010). Primacy effect or recency effect? A long-term memory test of Super Bowl commercials. *Journal of Consumer Behaviour*. *9 (1)*, 32–44.

Liu, J. S. (2001) *Monte Carlo Strategies in Scientific Computing*. NY: Springer-Verlag New York.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication research: An assessment and reporting of intercoder reliability. *Human Communication Research, 28,* 587–604.

Montgomery, D. C. (2008): *Design and Analysis of Experiments (7th Ed.)*, Hoboken, New Jersey: John Wiley & Sons.

Nan, X., &Faber, R. (2004). Advertising Theory: Reconceptualizing the Building Blocks. *Marketing Theory*, *4(1-2)*: 7-30.

Osgood, C. E. (1959). The representational model and relevant research methods. In I. de Sola Pool (ed.), *Trends in content analysis* (pp. 33–88). Urbana: U. of Illinois Press.

Perreault, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research, 26,* 135–148.

Popping, R. (1988). On agreement indices for nominal data. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research: Volume I, data collection and scaling* (pp. 90-105). New York: St. Martin's.

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27,* 258–284.

Riffe, D., Lacy, S., & Fico, F. G. (1998). *Analyzing media messages: Using quantitative content analysis in research.* Mahwah, N J: Lawrence Erlbaum Associates.

Riffe, D., Lacy, S., & Fico, F. G. (2005). *Analyzing media messages: Using quantitative content analysis in research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Rogot, E., & Goldberg I. D. (1966). A proposed index for measuring agreement in test-retest studies. *Journal of Chronic Diseases, 19(9)*, 991-1006

Rust, R. T., & Cooil, B. (1994). "Reliability measures for qualitative data: theory and implications." *Journal of Marketing Research*. 31(1), 1-14.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19,* 321–325.

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry, 44,* 172-177.

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry, 42,* 725–728.

Terry, W. S. (2005). Serial position effects in recall of television commercials. *Journal of General Psychology*, *132 (2)*, 151-163.

Thompson, W. D., & Walter, S. D. (1988). A reappraisal of the Kappa coefficient. *Journal of Clinical Epidemiology, 41 (10)*, 949-958.Thorson, E., & Zhao, X. (1997). Television viewing behavior as an index of commercial effectiveness. In W. D. Wells (ed.), Measuring advertising effectiveness (pp. 221-237). Hillsdale, New Jersey: Lawrence Erlbaum.

Vach, W. (2005). The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology, 58 (7)*, 655-661.

Zhao, X. (1989). *Effects of Commercial Position in Television Programming*. Doctoral Dissertation, University of Wisconsin – Madison.

Zhao, X. (1997). Clutter and serial order redefined and retested. *Journal of Advertising Research*, *37 (5),* 57-73.

Zhao, X. (2002a). A variable-based typology and a review of advertising-related persuasion research in 1990s. James Price Dillard and Michael Pfau (eds.): *The Persuasion Handbook: Developments in Theory and Practice,* pp. 495-512, Thousand Oaks, CA: Sage.

Zhao, X. (2002b). Partial causes and a typology of causal relations. *Journal of Jinan University, Social Science Edition*, 12:3, pp. 18-24.

Zhao, X. (2004a). Three types of information, three types of theories, three types of criteria, three types of truths (I). *Journal of Ocean University of China (Social Sciences Edition),* 65(3), pp. 29-36.

Zhao, X. (2004b). Three types of information, three types of theories, three types of criteria, three types of truths (II). *Journal of Ocean University of China (Social Sciences Edition),* 66(4), pp. 24-28.

Zhao, X. (2007a). A typology and probabilities of causal relations. *Journal of Ocean University of China (Social Sciences Edition),* 79:1, pp. 32-44.

Zhao, X. (2007b). Names, missions and constitution of journalism and mass communication —— a discussion with LI Xi-guang and PAN Zhong-dang. *Journal of Tsinghua University (Philosophy and Social Sciences)*, *22(5)*, 100-120

Zhao, X. (2011a, May). *When to use Cohen's* κ*, if ever?* Paper presented at the 61st annual conference of International Communication Association, Boston.

Zhao, X. (2011b, August). *When to use Scott's* π *or Krippendorff's* α*, if ever?* Paper presented at the 2011 annual conference of Association for Education in Journalism and Mass Communication, St. Louis.

Zhao, X., Deng, K., Feng, G. C., Zhu, L., & Chan, V. K. C. (2012a, May). Liberal-

conservative hierarchies for indices of inter-coder reliability. Paper presented at the 62nd annual conference of International Communication Association, Phoenix, Arizona.

Zhao, X., Liu, J. S., & Deng, K. (2012b). Assumptions behind inter-coder reliability indices. Charles T. Salmon (ed.) *Communication Yearbook 36*, 418-480. New York, NY: Routledge, Taylor & Francis Group.

Zhao, X; Shen, F, & Blake, K (1995). Position of TV advertisements in a natural Pod -- A preliminary analysis of concepts, measurements and effects. *Proceedings of the 1995 Conference of the American Academy of Advertising,* 154-161.

Zwick, R. (1988). Another look at inter-rater agreement. *Psychological Bulletin, 103,* 374–378.

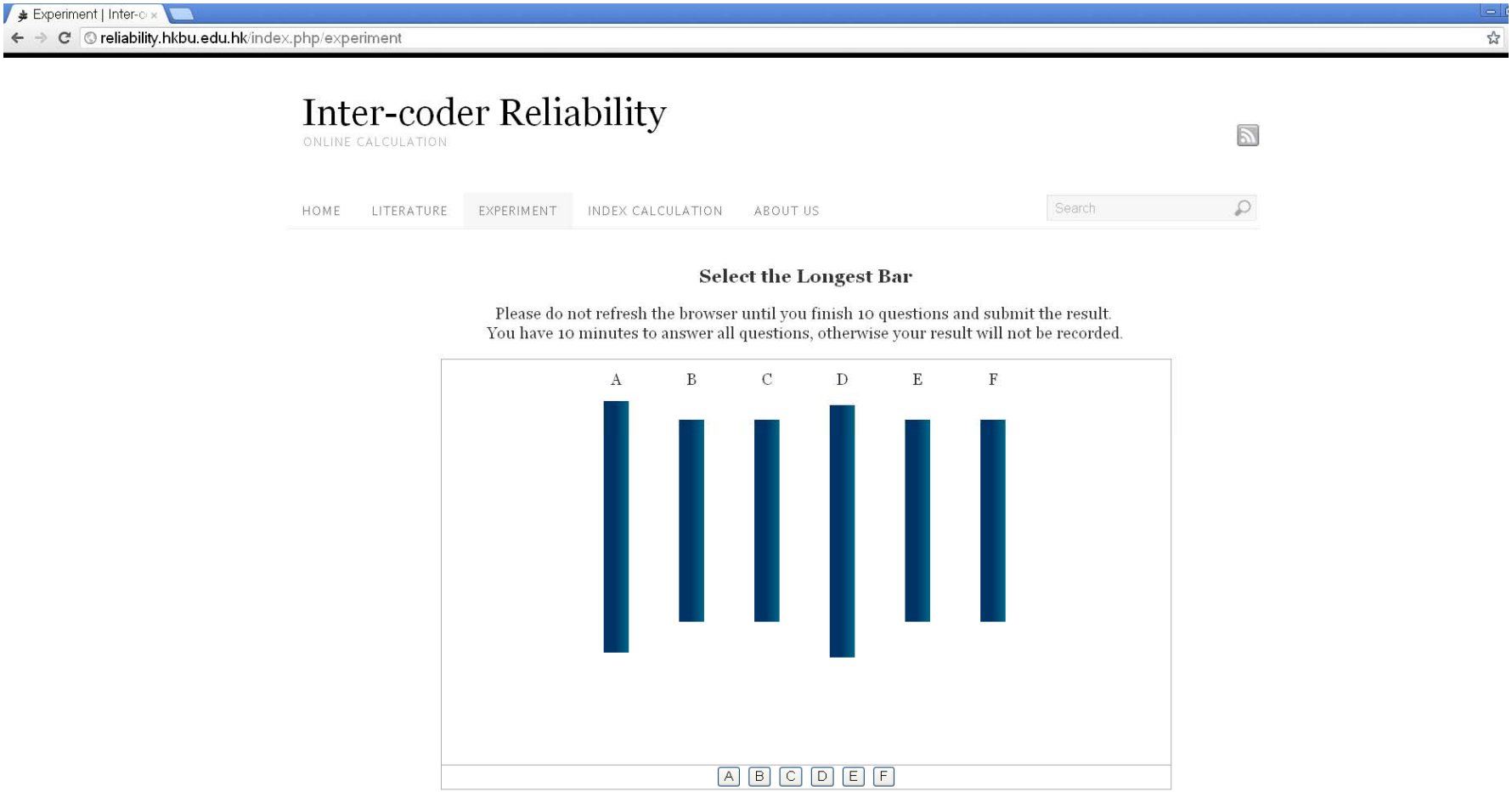

Figure 1. A sample screen seen by some coders in the BMC experiment (for category = 6, difficulty = 1).

Table 1

A Category (4) by Difficulty (8) by Distribution (3) Behavioral Monte Carlo (BMC) Experiment *

	Across: Distribution (Skew)		50%&50% (s_k=0.5)				25%&75% and 75%&25% (s_k=0.75)				1%&99% and 99%&1% (s_k=0.99)			
	Across: Category (K)		2	4	6	8	2	4	6	8	2	4	6	8
Difficulty (*df*)	difference in pixels (p_x)	Difficulty d_f=(8-p_x)/7												
	1	=1.000	4	4	4	4	4	4	4	4	4	4	4	4
	2	≈0.8571	4	4	4	4	4	4	4	4	4	4	4	4
	3	≈0.7143	4	4	4	4	4	4	4	4	4	4	4	4
	4	≈0.5714	4	4	4	4	4	4	4	4	4	4	4	4
	5	≈0.4286	4	4	4	4	4	4	4	4	4	4	4	4
	6	≈0.2857	4	4	4	4	4	4	4	4	4	4	4	4
	7	≈01429	4	4	4	4	4	4	4	4	4	4	4	4
	8	=0.0000	4	4	4	4	4	4	4	4	4	4	4	4

* Main cell entries are number of simulated coding sessions (subjects) in each experimental condition (cell).

Table 2

A Summary of Main Estimators and Estimands

Estimators Indices of intercoder reliability and their properties	Author(s) / Designer	*Estimator I* index (estimated true agreement)	*Estimator II* chance agreement estimated by each index
	Agreement Index (this manuscript)	a_i	ai_{ac}
	Bennett et al (1954)	S	S_{ac}
	Cohen (1960)	κ	κ_{ac}
	Gwet (2008, 2010)	AC_1	AC_{ac}
	Krippendorff (1970, 1980)	α	α_{ac}
	Percent Agreement (unknown author)	a_o	ao_{ac}
	Perrault and Leigh (1989)	I_r	Ir_{ac}
	Scott (1955)	π	π_{ac}
		Estimand I	*Estimand II*
Estimands Observed coder judgments	Main Estimands	a_t Observed true (non-chance) agreement	o_{ac} Observed chance agreement
	Secondary Estimands	a_r Observed right agreement	a_e Observed erroneous agreement
		a_o Observed Agreement	d_o Observed disagreement

Table 3

Main Findings from Behavioral Monte Carlo (BMC) Experiment (N_t=100, N_s=384)

		Manipulated Variables			Observed Variables		
		A. Category (K)	B. Target Distribution (Skew) (s_k)	C. Difficulty (d_f)	D. Percent Disagreement (d_o)	E. Chance Agreement (o_{ac})	F. True Agreement (a_t)
Group I. Manipulated variables	1. Category (K)	**1.000*** **	**.000**	**.000**	**.044**	**-.138****	**.059**
	2. Distribution (Skew) (s_k)	**.000**	**1.000*** **	**.000**	**-.004**	**-.023**	**.016**
	3. Difficulty (d_f)	**.000**	**.000**	**1.000*** **	**.882*** **	**.765*** **	**-.880*** **
Group II. Chance Agreement Estimated by Indices	4. Percent Agreement (ao_{ac}=0)	**---**	**---**	**---**	**---**	**---**	---
	5. Gwet (AC_{ac})	**-.813*** **	**-.197*** **	**.095**	**.088**	**.273*** **	-.202***
	6. Bennett et al (S_{ac})	**-.929*** **	**.000**	**.000**	**-.065**	**.146****	-.053
	7. Perrault and Leigh (Ir_{ac})	**-.929*** **	**.000**	**.000**	**-.065**	**.146****	-.202***
	8. Cohen (κ_{ac})	**-.117***	**.659*** **	**-.354*** **	**-.475*** **	**-.390*** **	.461***
	9. Scott (π_{ac})	**-.116***	**.661*** **	**-.351*** **	**-.472*** **	**-.388*** **	.458***
	10. Krippendorff (α_{ac})	**-.116***	**.661*** **	**-.351*** **	**-.472*** **	**-.388*** **	.458***
	11. Agreement Index (ai_{ac})	**-.186*** **	**-.008**	**.853*** **	**.946*** **	**.745*** **	-.900***
Group III. Inter-coder Reliability Indices	12. Percent Agreement a_o	-.044	.004	-.882***	-1.000***	-.729***	**.917*** **
	13. Gwet's AC_1	.351***	.057	-.744***	-.844***	-.743***	**.849*** **
	14. Bennett et al's S	.418***	.003	-.752***	-.832***	-.723***	**.831*** **
	15. Perrault and Leigh's Ir	.430***	-.021	-.659***	-.747***	-.697***	**.774*** **
	16. Cohen's κ	.038	-.540***	-.624***	-.644***	-.415***	**.559*** **
	17. Scott's π	.037	-.541***	-.624***	-.644***	-.415***	**.559*** **
	18. Krippendorff's α	.037	-.541***	-.624***	-.644***	-.415***	**.559*** **
	19. Agreement Index (a_i)	.071	.006	-.879***	-.987***	-.747***	**.921*** **

ao_{ac} is the chance agreement estimated by percent agreement a_o. Because it is a constant at zero, the correlation coefficients cannot be calculated. The line serves as a reminder that a_o has an estimated chance agreement.

Table 4

A Typology of Coding Based on Black-White Randomness Assumption

	Systematic Coding		Chance Coding
Right: Source of Variation Down: Coding Outcome	1. Undesired Miscoding	2. Honest & Accurate Coding	3. Honest & Radom Coding
Agreements	A_v: Agreement from Deliberate Miscoding (Undesired and assumed zero for data analysis) Systematic miscoding, such as deliberate quota, which should be prevented through monitoring and training.	A_a: True (Accurate) Agreement (Desired) Most desirable.	A_c: Chance Agreement (Undesired and needs to be estimated) Undesirable but removable in estimation.
Disagreements	D_v: Disagreement from Deliberate Miscoding (Undesired and assumed zero for data analysis) Systematic miscoding, such as deliberate quota, which should be prevented through monitoring and training.	D_a: Accurate Disagreement This concept does not exist, hence the cell is by definition empty.	D_c Chance Disagreement (Undesired and observed) Undesirable but removable in estimation, and useful in estimating the systematic disagreements.

Table 5

Agreements and Disagreements with Two Coders and Three Categories

		Coder 1					
		Category 1	Category 2	Category 3	D_{2c}	d_{2c}	N_{2c}
Coder 2	Category 1	J_{11}	J_{21}	J_{31}	$D_{21}=J_{21}+J_{31}$	$d_{21}=D_{21}/D_o$	$N_{21}=J_{11}+J_{21}+J_{31}$
	Category 2	J_{12}	J_{22}	J_{32}	$D_{22}=J_{12}+J_{32}$	$d_{22}=D_{22}/D_o$	$N_{22}=J_{12}+J_{22}+J_{32}$
	Category 3	J_{13}	J_{23}	J_{33}	$D_{23}=J_{13}+J_{23}$	$d_{23}=D_{23}/D_o$	$N_{32}=J_{13}+J_{23}+J_{33}$
	D_{1c}	$D_{11}=J_{12}+J_{13}$	$D_{12}=J_{21}+J_{23}$	$D_{13}=J_{31}+J_{32}$	$D_o=D_{11}+D_{12}+D_{13}$ $=D_{21}+D_{22}+D_{23}$		
	d_{1c}	$d_{11}=D_{11}/D_o$	$d_{12}=D_{12}/D_o$	$d_{13}=D_{13}/D_o$		$d_o= D_o/N$	
	N_{1c}	$N_{11}=J_{11}+J_{12}+J_{13}$	$N_{12}=J_{21}+J_{22}+J_{23}$	$N_{13}=J_{31}+J_{32}+J_{33}$			$N=N_{11}+N_{12}+N_{13}$ $=N_{21}+N_{22}+N_{23}$
	$a_o=(J_{11}+J_{22}+J_{33})/N$		$c_c= d_{11}*d_{21}+ d_{12}*d_{22}+ d_{13}*d_{23}$		$a_c= d_o*(c_c/(1-c_c)$		$a_i=a_o-a_c$

Table 6

A Typology of Coding Based on Mixed-Randomness Assumption

	Systematic Coding		Chance Affected Coding	
Right: Source of Variation Down: Coding Outcome	1. Deliberate Miscoding	2. Honest and Accurate Coding	3. Honest and Mixed Coding	4. Honest and Pure Chance Coding
Agreements	a_v (Deliberately miscoded agreements, Undesired and assumed zero for data analysis) Systematic miscoding, such as deliberate quota, which should be prevented through monitoring and training.	a_a (Purely Accurate Agreements, Desired) Most desirable.	a_m (Mixed chance agreement. Non-consequential) At least one coder codes systematically, hence accurately, while at least another codes randomly. It's not bad because the agreed result is still accurate	a_c (Pure Chance Agreement, undesired and needs to be estimated) Undesirable but removable in estimation. $a_c=a_n+a_e$ a_n: correct agreement that is purely by chance a_e: erroneous agreement, which is assumed to be all purely by chance
Disagreements	d_v (Deliberately miscoded disagreements, undesired and assumed zero for data analysis) Systematic miscoding, such as deliberate quota, which should be prevented through monitoring and training.	d_a: True, Purely Accurate Disagreement This sub-concept does not exist, hence the cell is by definition empty.	d_m (Mixed Disagreement, Undesired and observed) Undesirable but observed together with D_c, and useful in estimating the pure-chance agreements.	d_c (Pure Chance Disagreement, undesired and observed) Undesirable but observed together with D_m, and useful in estimating the pure-chance agreements .

1. I call d_c "pure chance disagreement," or "chance disagreement."
2. $d_f=d_m+d_c$. I call d_f "chance affected disagreement" or "chance disagreement."
3. Because d_a is by definition zero and d_v is assumed zero, $d_m+d_c=d_f$ constitutes all observed disagreement d_o, that is, $d_f=d_o$. Hence all disagreements are chance disagreements, while all agreements are not chance agreements.
4. I call a_c "pure chance agreement" or "chance agreement."
5. $a_f=a_m+a_c$. I call a_f "chance affected agreement" or "chance agreement."
6. $a_r=a_a+a_m+a_n$. I call a_r "right agreement." Also, $a_r=a_o-a_e$.
7. $a_t=a_a+a_m$. I call a_t "true agreement."
8. $c_q=a_c+d_m+d_c=c_c+d_o$ where c_q is defined as "consequential chance coding."
9. Note the difference between a_a, a_t, a_r, and a_n. They are all "correct." But their sources and/or compositions are different.
10. The task is to estimate a_f.

Table 7

Chance Agreements with Three Categories

Across: Coder A Down: Coder B	a. Random between 3 categories	b. Random between Categories 1&2	c. Radom between Categories 2&3	d. Accurate Coding
1. Random between 3 categories	a1. Pure Chance Agreement (Correct & Erroneous)	b1. Pure Chance Agreement (Correct & Erroneous)	c1. Pure Chance Agreement (Correct & Erroneous)	d1. Mixed Chance Agreement (all correct)
2. Random between Categories 1&2	a2. Pure Chance Agreement (Correct & Erroneous)	b2. Pure Chance Agreement (Correct & Erroneous)	c2. Pure Chance Agreement (Correct & Erroneous)	d2. Mixed Chance Agreement (all correct)
3. Random between Categories 2&3	a3. Pure Chance Agreement (Correct & Erroneous)	b3. Pure Chance Agreement (Correct & Erroneous)	c3. Pure Chance Agreement (Correct & Erroneous)	d3. Mixed Chance Agreement (all correct)
4. Accurate Coding	a4. Mixed Chance Agreement (all correct)	b4. Mixed Chance Agreement (all correct)	c4. Mixed Chance Agreement (all correct)	d4. Non-Chance Agreement

Table 8

A Typology of Chance-Affected Agreements

Across: Type of Coding Down: Resulted Agreements	Mixed Coding a_m+d_m	Pure Chance Coding $a_c+d_c=a_n+a_e+d_c$
Correct Chance Agreement ($a_r=a_n+a_m$)	mixed chance agreement (a_m)	correct chance agreement (a_n)
Erroneous Chance Agreement (a_e)	*	erroneous chance agreement (a_e)

* This is an empty cell, as mixed coding cannot produce erroneous agreements.

Table 9

An Example for Calculating Agreement Index a_i with Two Coders and Three Categories

		Coder 1					
		Category 1	Category 2	Category 3	D_{2c}	d_{2c}	N_{2c}
Coder 2	Category 1	J_{11} =3	J_{21} =5	J_{31} =1	D_{21} =5+1=6	d_{21} =6/27=0.222	N_{21} =3+5+7=9
	Category 2	J_{12} =6	J_{22} =8	J_{32} =4	D_{22} =6+4=10	d_{22} =10/27=0.370	N_{22} =6+8+4=18
	Category 3	J_{13} =2	J_{23} =9	J_{33} =7	D_{23} =2+9=11	d_{23} =11/27=.407	N_{32} =2+9+7=18
	D_{1c}	D_{11} =6+2=8	D_{12} 5+9=14	D_{13}= 1+4=5	D_o= 6+10+11=27		
	d_{1c}	d_{11} =8/27=.296	d_{12} =14/27=.519	d_{13} =5/27=.185		d_o =27/45=0.6	
	N_{1c}	N_{11} =3+6+2=11	N_{12} =5+8+9=22	N_{13} =1+4+7=12			N =9+18+18=45
	a_o =0.4	c_c =.296*.222+.519*.370+.185*.407=0.333			a_c =.6*.333/(1-.333)=0.3		a_i =.4-.3=0.1