CONSENSUS ANALYSIS'S UNDISCUSSED SAMPLING ISSUE: HOW MANY QUESTIONS ARE NEEDED TO ESTABLISH CREDIBLE ASSESSMENTS OF RESPONDENT-BY-RESPONDENT SIMILARITY?

John Gatewood
Lehigh University

SASci / SfAA Meetings in Albuquerque, N.M., March 2014

ABSTRACT

Consensus analysis rests upon people's responses to batteries of forced-choice questions. Two sampling issues are involved in such data collections. The first concerns respondents, and there are well-known ways to select respondents that ensure findings can be generalized to larger populations. The second sampling issue is more subtle – formulating a battery of questions that adequately samples respondents' knowledge. More specifically, how many questions are needed to establish credible respondent-by-respondent similarity measures (which are what consensus analysis actually analyzes)? This paper discusses different approaches to this 'N of questions' issue, two based on general statistical reasoning and one based on simulations.

PREVIEW

1. The Problem

   How many questions are needed to establish credible assessments of respondent-by-respondent similarity?

2. Conceptual Framework for Solving Problem
3. Three 'Solutions'
   A. Simplest mathematical approach (most conservative)
   B. Complicated mathematical approach (yet-to-be-developed)
   C. Brute force approach (simulations)
4. Johnny's Advice

1. THE "N OF QUESTIONS" PROBLEM

- Cultural Consensus Analysis (CCA) estimates the degree of "shared" knowledge based on respondents' answers to a battery of fixed-format questions.
- Respondent-by-Item data is converted to a Respondent-by-Respondent matrix in which the cell-values are pairwise measures of similarity:
  -- PERCENTAGE OF MATCHES … if questions involve categorical answers, or
  -- PEARSON R … if questions involve ratings or rankings.
- The R x R similarity matrix is then factor analyzed (after correcting-for-guessing if similarity is Match%) via a least squares method, i.e., minimum residual factoring.
- Key indicators: (a) mean 1st factor loading ≥ .50, (b) ratio of 1st to 2nd eigenvalues ≥ 4.0, and (c) few negative 1st factor loadings.
- But, **HOW MANY QUESTIONS ARE NEEDED** to establish credible assessments of similarity among the respondents?

(We'll presume all the questions are GOOD ones and the battery of questions is COUNTER-BALANCED.)

NOTE: Susan Weller's rule-of-thumb advice concerning "N of Questions" …

"Can I do a consensus analysis on responses to only four questions? Although this can be done, it is not advisable. All the methods described above rely on the agreement between people across questions (questions are the unit of analysis). To estimate the agreement between each pair of informants, a greater number of questions provides a more stable (and thus, better) estimate. **At least twenty questions** are recommended to obtain reasonable estimates" (Weller, 2007: 350).

Weller, Susan C. 2007. Cultural consensus theory: Applications and frequently asked questions. *Field Methods* 19: 339-368.

2. CONCEPTUAL FRAMEWORK FOR SOLUTION

- There is a fairly simple relation between R x R measures of similarity and CCA's 1st factor loadings:

$$r_{ij} = r_{iT} \cdot r_{jT} \approx D_i \cdot D_j$$

… i.e., the observed correlation between Person i and Person j is assumed to be equal to the product of their, respective, correlations with the Truth (i.e., the 'Answer Key').

- **NET EFFECT:**
  The mean r (or, the mean chance-corrected Match%) in the R x R similarity matrix needs to be ≥ .25 in order for the mean 1st factor loading from CCA to be ≥ .50 .

- **"N OF QUESTIONS" PROBLEM (re-phrased):**
  How many questions are needed for an observed similarity value of .25 or greater to be "statistically significant" … as unlikely to have happened just by chance?

3. THREE 'SOLUTIONS'

A. SIMPLEST MATHEMATICAL APPROACH …
Use known probability distributions to calculate number of questions needed for a Pearson r (or Match%) ≥ .25 between a single pair of respondents to be statistically significant.

B. COMPLICATED MATHEMATICAL APPROACH …
Two-stage probability calculation involving both "N of questions" and "N of pairwise comparisons" to assess statistical significance of an R x R matrix having a ***mean*** Pearson r (or ***mean*** Match%) ≥ .25 .

C. BRUTE FORCE (SIMULATION) APPROACH …
Generate multiple 30 Respondent x 30 Question data files using the "know-or-guess" process model. Do CCA on only the first 6 questions, then the first 12, then the first 18, then the first 24, and then all 30 questions. Examine results to identify the approximate "N of Questions" at which consensus indicators seem to stabilize.

3-A. SIMPLEST APPROACH

- For categorical response questions (true/false, multiple-choice), use **BINOMIAL DISTRIBUTIONS** to determine number of questions needed to make a corrected-for-guessing Match% = .25 statistically significant.

- For scalar response questions (ratings, rankings), use **t-DISTRIBUTIONS** to determine number of questions needed to make a Pearson r = .25 statistically significant.

* * * go to "N of Questions Calculator" spreadsheet * * *

Microsoft Excel Worksheet

- **CALCULATED ANSWERS** … for N of Questions required to make a single pairwise measure of similarity ≥ .25 statistically significant (alpha = .05, one-tailed):
- Categorical data → Match% is measure of similarity
  - with 2 categories … minimum of 53 questions
  - with 3 categories … minimum of 30 questions
  - with 4 categories … minimum of 23 questions
  - with 5 categories … minimum of 20 questions
  - with 6 categories … minimum of 16 questions
- Scalar data → Pearson r is measure of similarity
  - ratings / rankings … minimum of 45 questions

3-B. COMPLICATED (CORRECT) APPROACH

- The "Simplest" Approach is appropriate for a single comparison of two people, but it is surely too conservative when applied to the ***MEAN*** of multiple pairwise comparisons.

- The conceptually-correct mathematical approach, however, is "yet-to-be-developed."

  Good news: I have sweet-talked a mathematician colleague to work on it.

  Bad news: He says it is a "very difficult" problem.

- So … today, I'll just note the intuitions underlying this "Complicated" Approach and sketch the mathematical problems awaiting solutions.

- **BASIC INTUITION:** Each cell in an R x R similarity matrix rests on the same "N of Questions," but the matrix contains many quasi-independent pairwise comparisons. Surely, then, the average of the off-diagonal cells is a more reliable estimate of the true population parameter than any single cell, because the average would be less susceptible to sampling error.

  In short, the statistical significance of the **mean** Pearson r in an R x R matrix will involve two N's… N of Questions and N of Respondent-Pairings.

- **GOAL OF APPROACH:** Find a way to calculate N of Questions based on N of Respondent-Pairings, because there's a trade-off between the two. And, following this logic, the same degree of statistical significance could be achieved with fewer questions than called for by the "Simplest" Approach.

But, there are some "very difficult" mathematical problems that must be solved before this Complicated Approach can be realized … (next slide)

- **Math Question #**1:
  Say you have several studies involving the same two variables but with independent samples (and let's say of the same N just for simplicity). It's relatively easy to calculate the mean Pearson r of these studies (e.g., using Fisher's r-to-z transformation then back again), but how do you calculate the probability of that *mean* Pearson r ?

- **Math Question #2**:
  How can one assess the "statistical significance" for the *mean* Pearson r of the off-diagonal values in an R x R correlation matrix, i.e., the average of non-redundant cells, each of which represents a correlation based on the same N of Questions?
  (Here, of course, the 'different' correlation coefficients are not completely independent, e.g., a 10 person x 10 person correlation matrix has 10 "independent people" but 45 different pairings in the lower-half of the matrix.)

... AWAITING ANSWERS TO THESE QUESTIONS ... from a mathematician.

3-C. BRUTE FORCE APPROACH

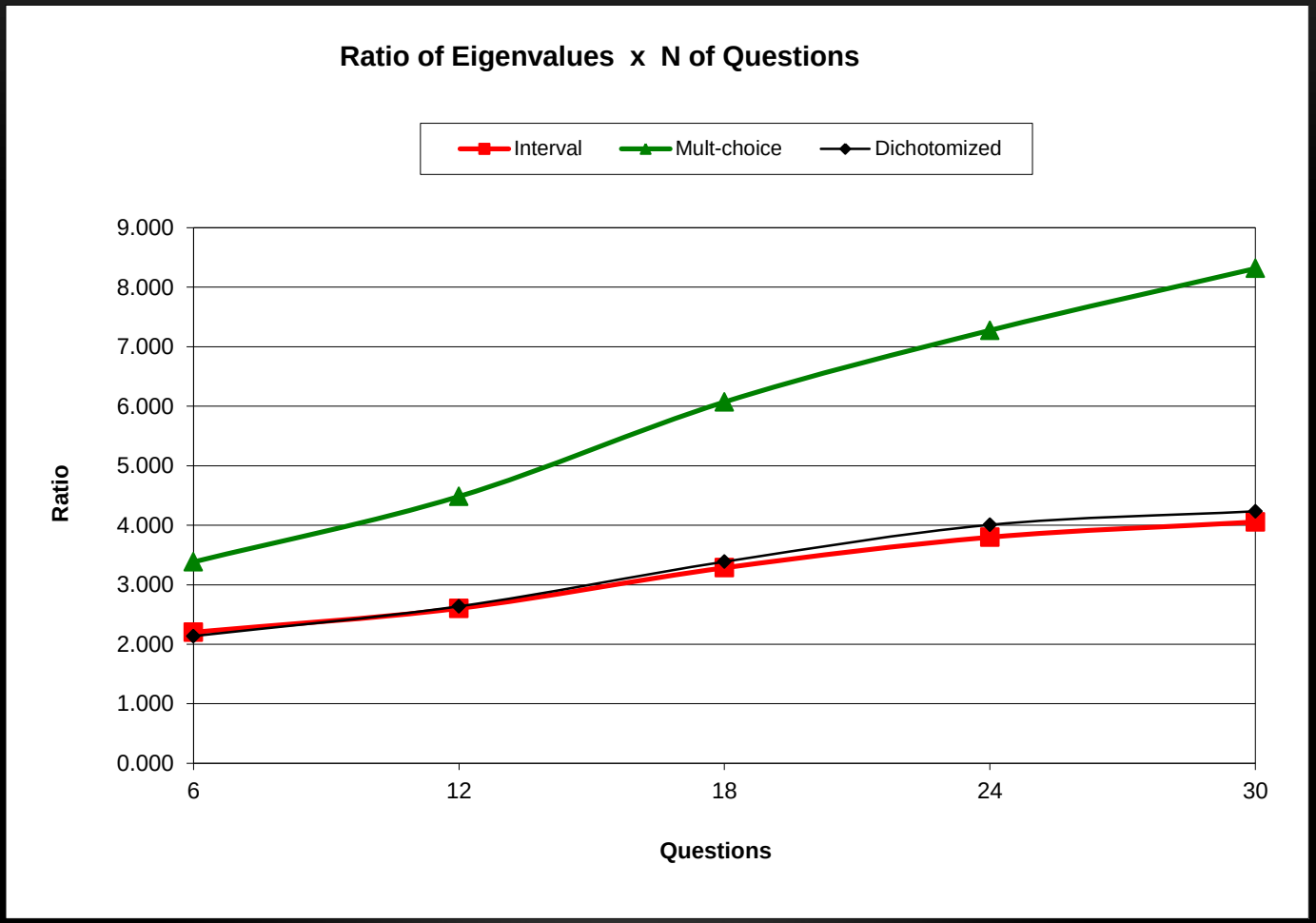
- Varying "N of Questions" in ***simulated data*** reveals interesting effects vis-à-vis CCA.
- Ten data files were generated with an Excel spreadsheet using the same parameters:
  1. The initial files represent ten 'runs' of 30 Respondents answering 30 Questions.
  2. Each question has a fixed correct answer – an integer from 1 to 6 – and the Answer Key is simply a repeated sequence of 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, ...; hence, for each subset of six consecutive questions, there are equal numbers of 1s, 2s, 3s, etc., in the Answer Key.
  3. For each question, a respondent either "knows" its correct answer or "guesses"; the probability of knowing is set at P(Know) = .50 for each respondent for each question; and when guessing, respondents do so equiprobably.

(Happy to show the GENERATING SPREADSHEET to anyone who's interested.)

- Each of the initial 30 R x 30 Q data files produced five files with different numbers of questions, as follows:
  - ten 'runs' of 30 R x **6 Q** … the first six questions in initial data file
  - ten 'runs' of 30 R x **12 Q** … the first twelve questions
  - ten 'runs' of 30 R x **18 Q** … the first eighteen questions
  - ten 'runs' of 30 R x **24 Q** … the first twenty-four questions
  - ten 'runs' of 30 R x **30 Q** … all thirty questions
- Each of these 50 data files was analyzed three ways using Anthropac 4.983X:
  - Interval Method … 1-to-6 response scale
  - Multiple-choice Method … 6 response categories {1,2,3,4,5,6}
  - Multiple-choice Method … dichotomized responses {1,2,3} → 0 and {4,5,6} → 1

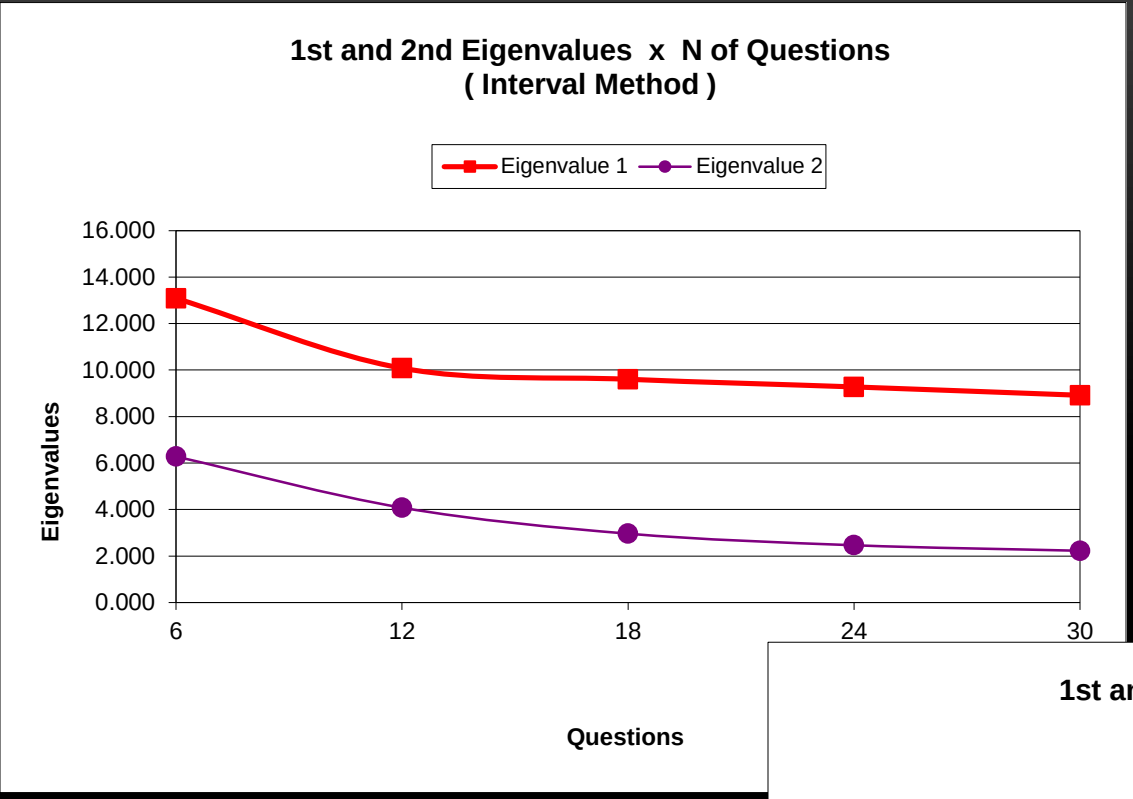
- Both the **HANDOUT** and the graphs you'll see in a minute show averages over ten separate simulation 'runs' for each of five different-length batteries of questions.

- Key thing to look for:
  How CCA-relevant variables (the vertical axes) change as the "N of Questions" increases (the horizontal axis).



-- RATIO of eigenvalues increases almost linearly with “N of Questions.”

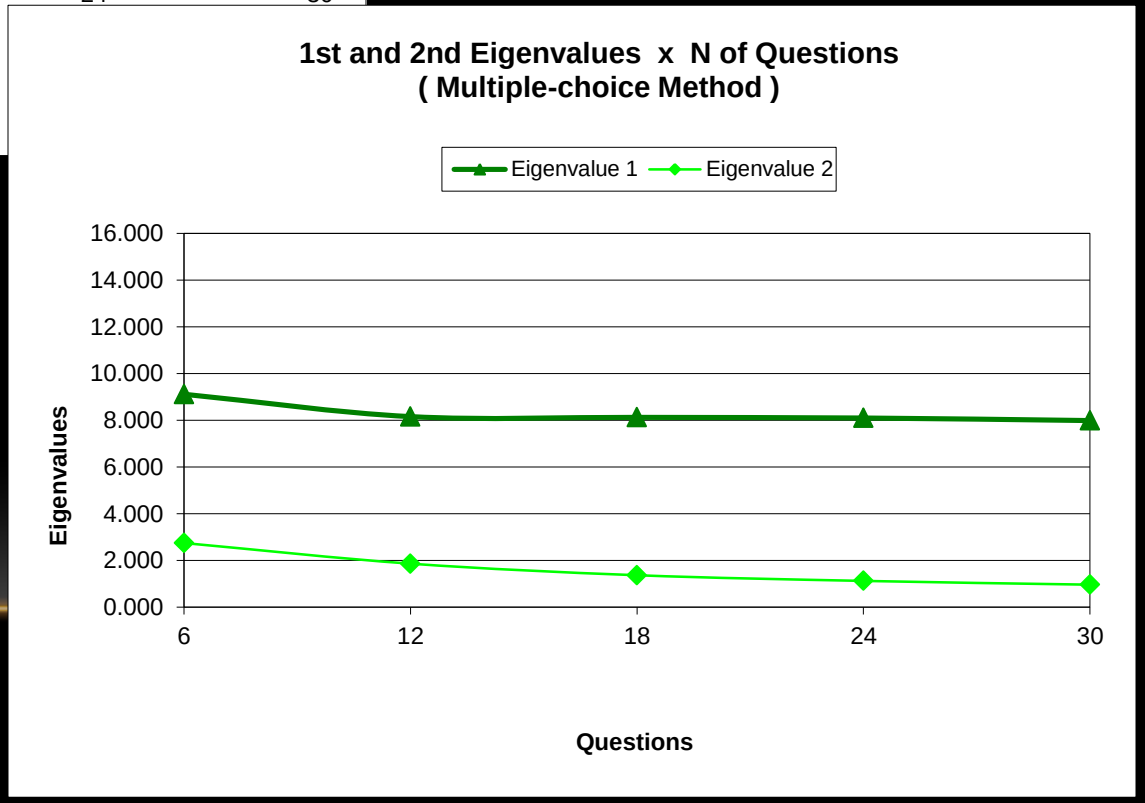
-- Multiple-choice CCA of dichotomized data ≈ Interval CCA.

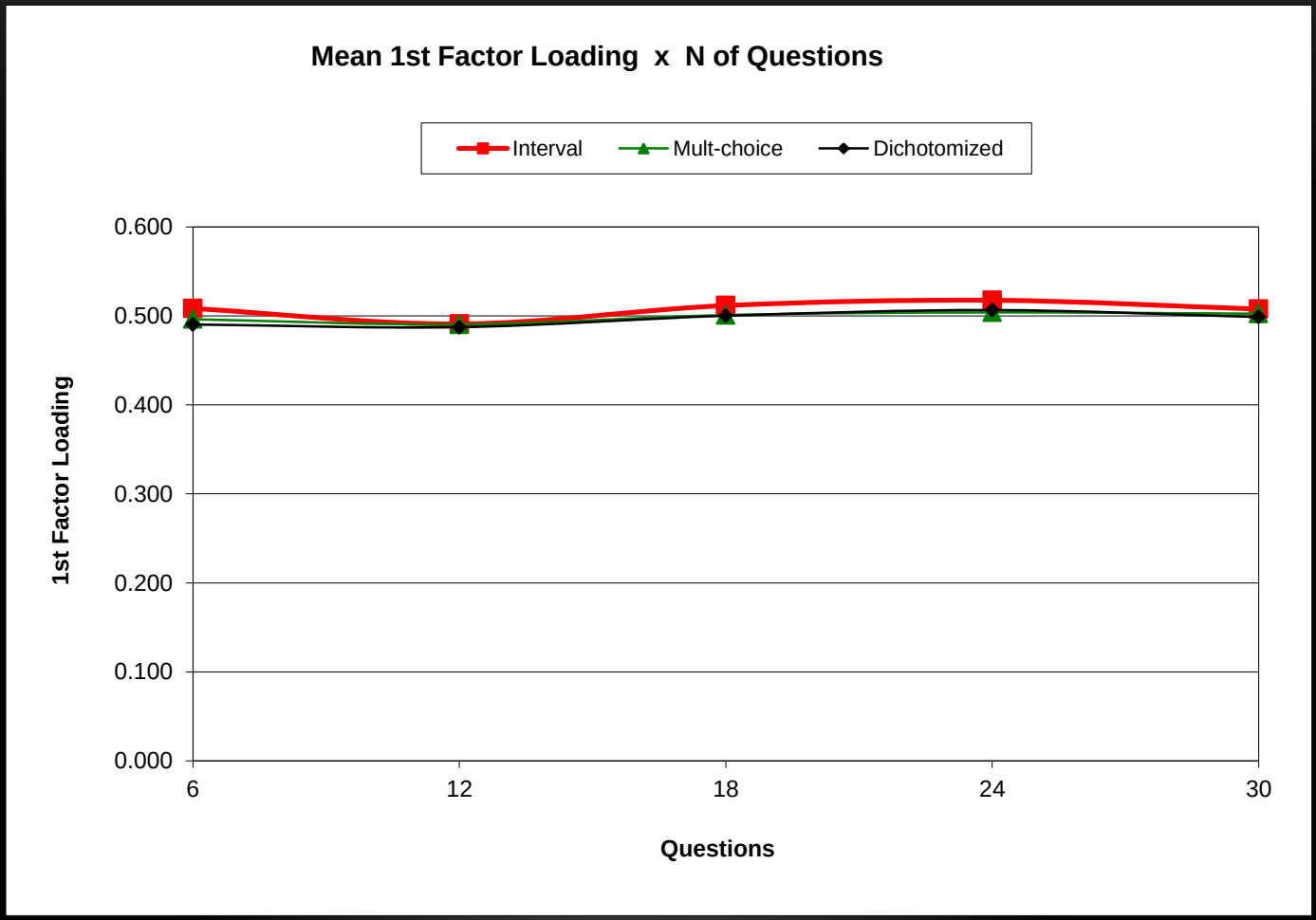


-- While RATIO increases with "N of Questions" (previous slide), the **eigenvalues themselves decrease** as "N of Questions" increases.

-- The almost linear increase in RATIO is due mostly to faster rate of decrease in 2nd Factor's eigenvalue.
(because 'randomness' of guessing becomes more clearly random)

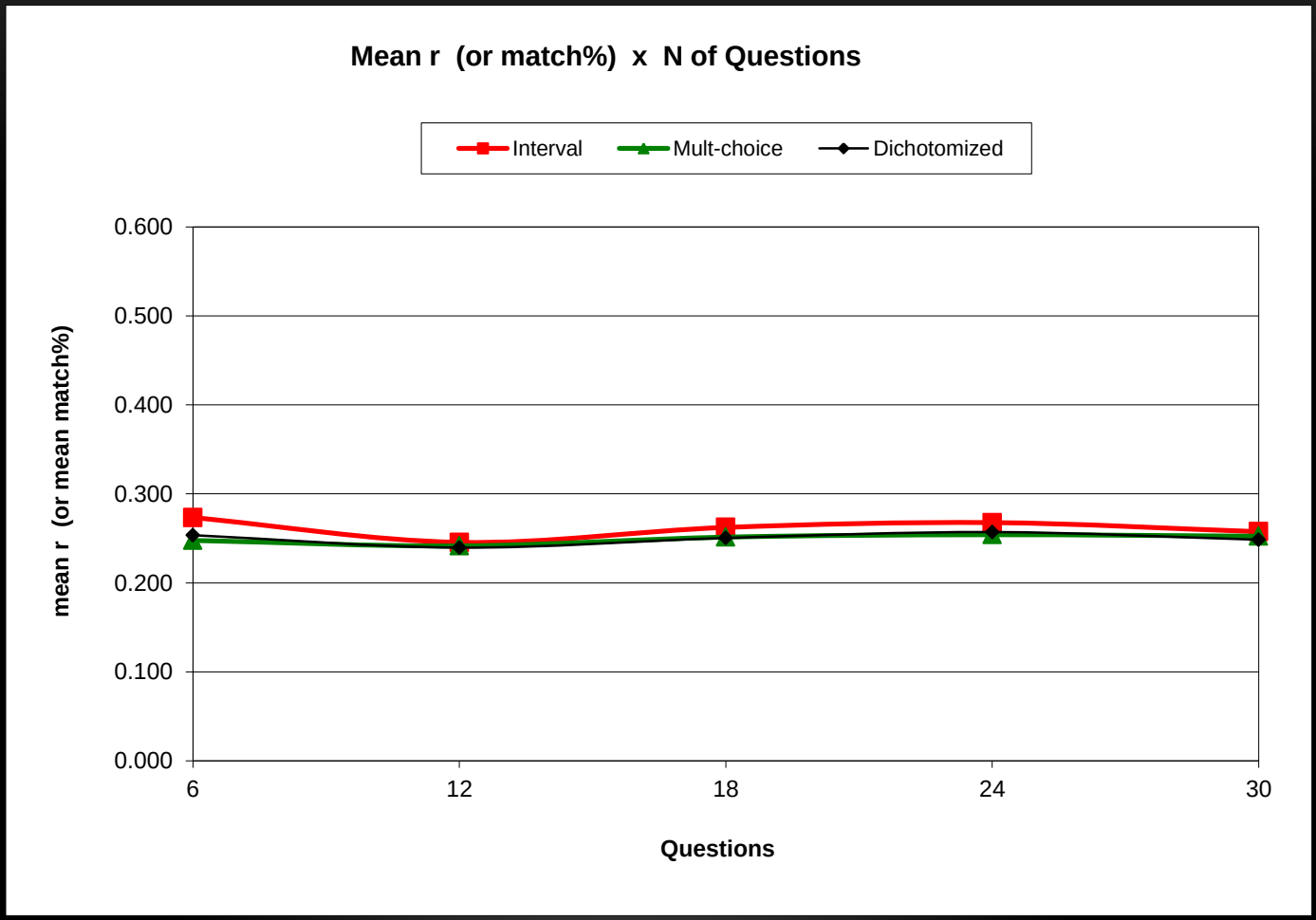
… above observations are true for both Interval & Multiple-choice CCA.



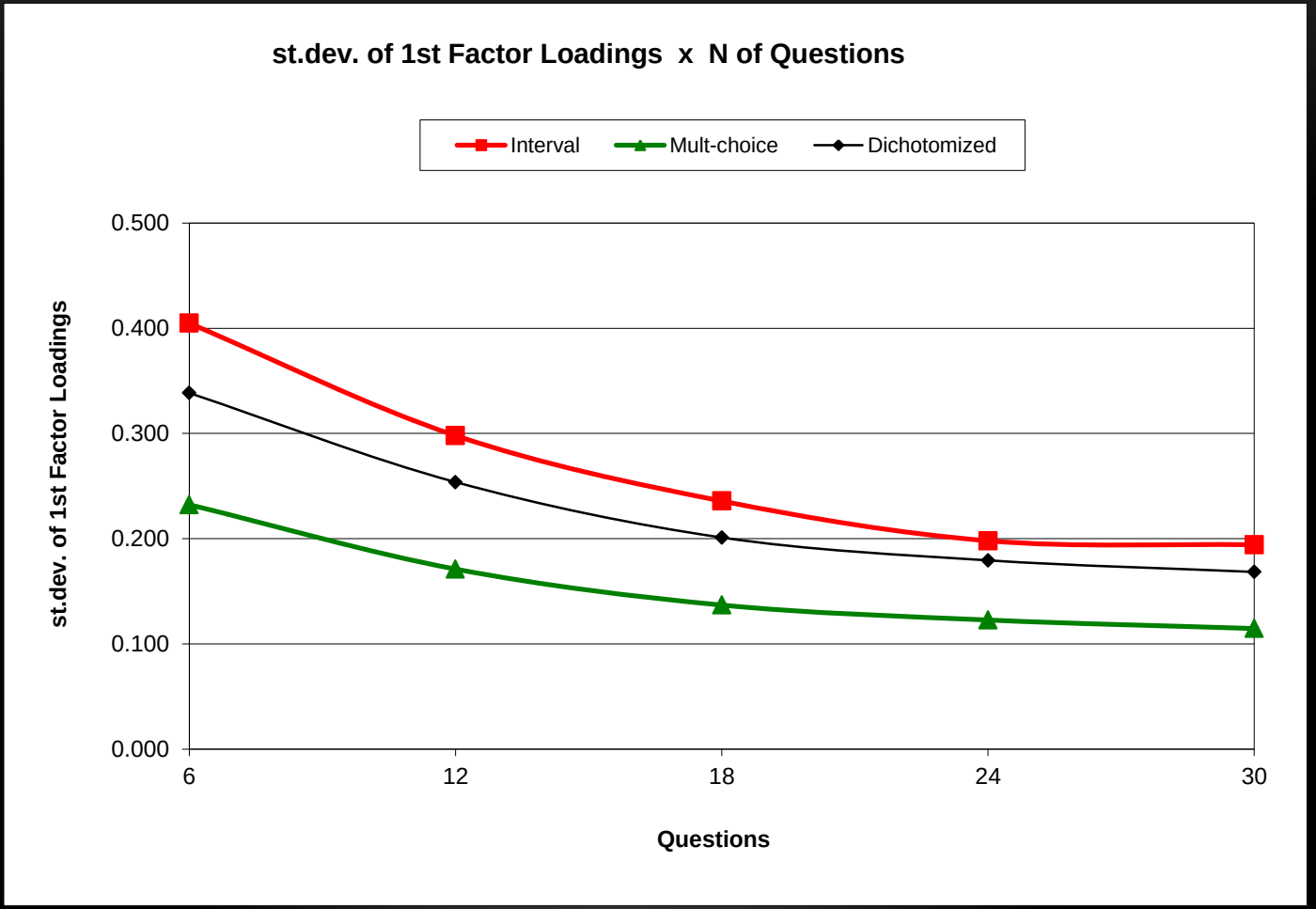


-- "N of Questions" has virtually **no effect** on the MEAN 1st Factor loading.

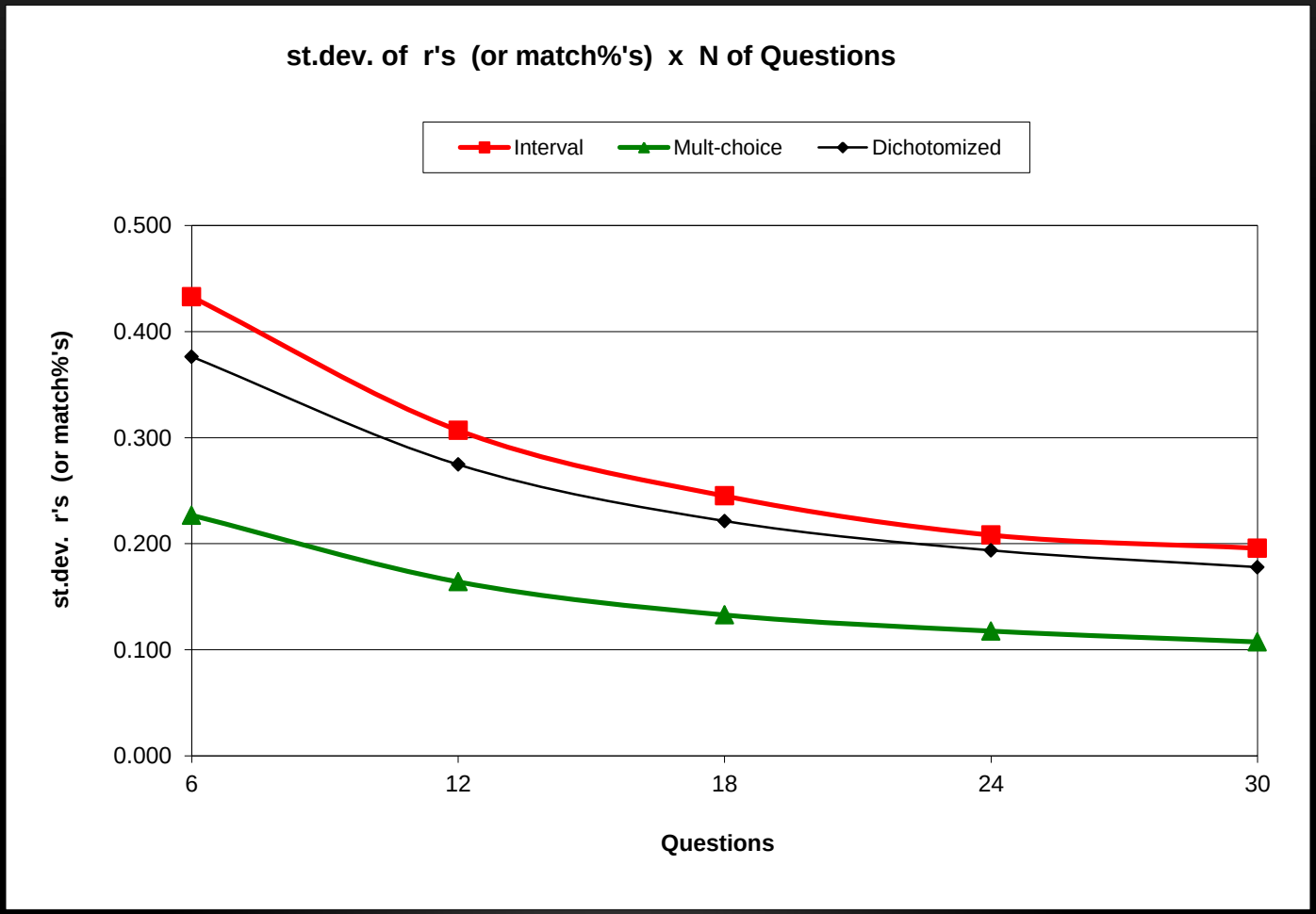
-- All three CCA Methods do excellent job of recovering built-in knowledge level.



-- “N of Questions” has virtually **no effect** on the MEAN measures of R x R similarity, which is why no effect on mean 1st Factor loading (previous slide).

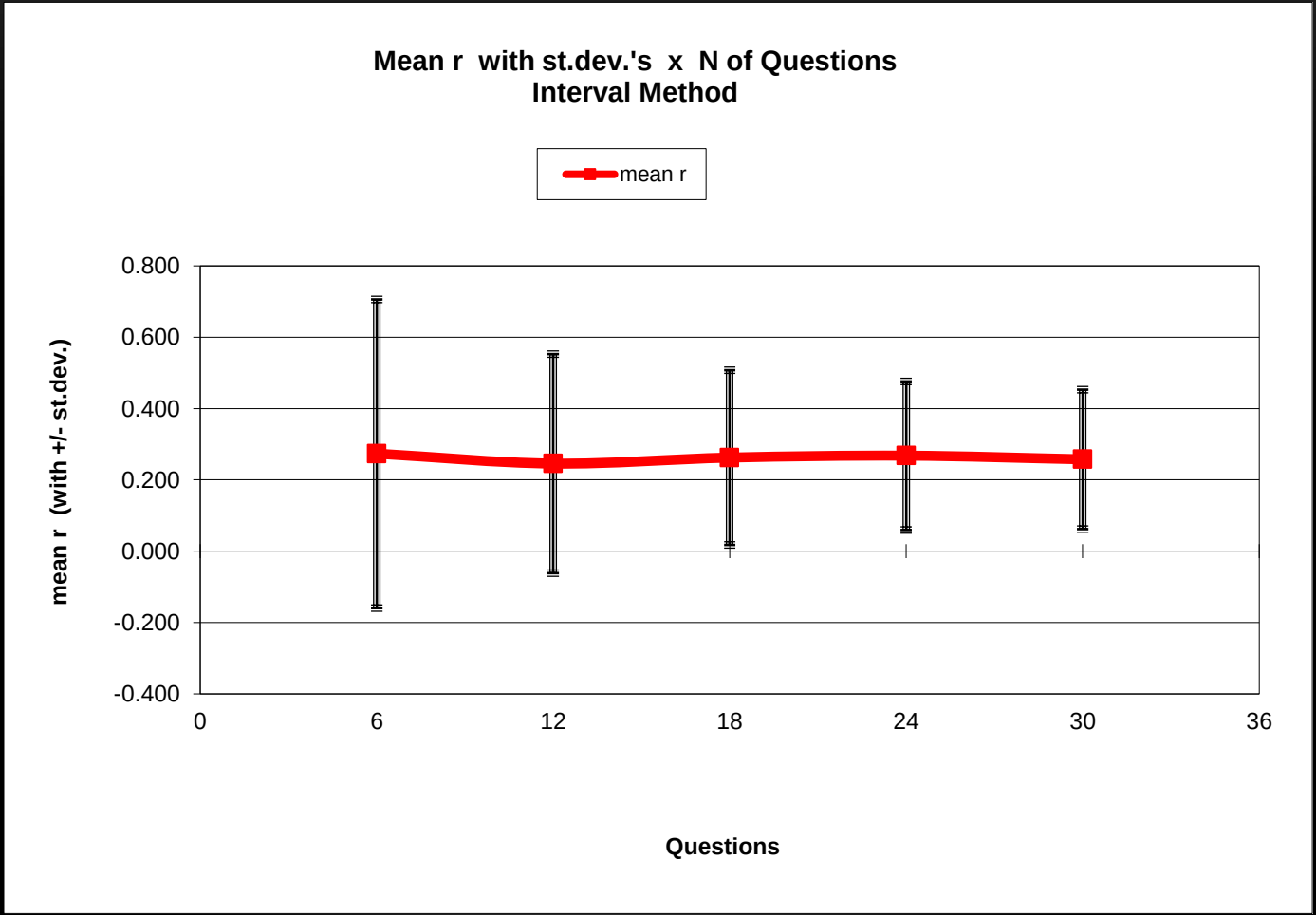


-- "N of Questions" has a **strong non-linear effect** on standard deviation of 1st Factor loadings … more questions → less dispersion.



-- "N of Questions" has a **strong non-linear effect** on standard deviation of measures of R x R similarity … more questions → less dispersion.

MAIN OBSERVATION: ***Dispersions*** about the MEAN measures of R x R similarity seem to "stabilize" (slope gets flatter and flatter) as N of Questions goes from 24 to 30.



BOTTOM LINE: CCA needs at least ~ 30 questions to sort out "signal" from "noise."

CONCLUSIONS FROM SIMULATIONS:

1. RATIO of eigenvalues is a volatile indicator of consensus. It is susceptible not only to "distributional pattern" of knowledge [previous SASci talk], but also to "N of Questions."
2. By contrast, the MEAN 1st Factor loading (a.k.a., the average shared knowledge) is quite stable and well-"recovered" by CCA. This indicator of consensus – rather than the Ratio – is the most important and should be the first-reported.
3. There is a reason "N of Questions" has its observed effects – on the Ratio and magnitudes of the eigenvalues, on the standard deviations of 1st Factor loadings, and on the standard deviations of the basic measures of R x R similarity. The reason is:
   -- Few questions can produce spurious 'patterns' just by chance.
   -- With more questions, such spurious patterns melt away into randomness.
   -- ***In short, more questions improves the ability to detect "signal" from "noise."***
4. **So, how many questions are needed for 'credible' CCA?**

   ANSWER … ~ 30 would be a good rule-of-thumb
   With 30 questions, all three criteria of consensus (ratio, mean 1st factor loading, few negatives) were met … *as they should have been given the simulation settings.*

4. JOHNNY'S ADVICE

A. More questions are definitely gooder than fewer … (and generally more than 20).

B. If you want to be very conservative, then you'll need at least:
   a. 45 rating questions or items for ranking task;
   b. 53 dichotomous-response questions;
   c. And, for "multiple-choice" with >2 response options:
      30 questions with 3 options
      23 questions with 4 options
      20 questions with 5 options
      16 questions with 6 options.

C. Alternatively, if you think my liberal interpretation of the brute force simulations holds water, then maybe you can **get away with as few as ~ 30 questions**, but not less.

Lastly, when using Anthropac's "Interval" method for rating data or "Covariance" method for dichotomous data, make sure your items are roughly counter-balanced, either through data collection using paired-opposite phrasings of questions or ex post facto by re-polarizing some items.
(Don't worry about this if data are based on rankings or q-sorts, or when using "Multiple-choice" method.)

THANK YOU !

Questions ? … Comments ?