RASCH MEASUREMENT TRANSACTIONS PARTS 1 and 2

Edited by

John M. Linacre
UNIVERSITY OF CHICAGO

MESA

RASCH MEASUREMENT TRANSACTIONS
PART 1

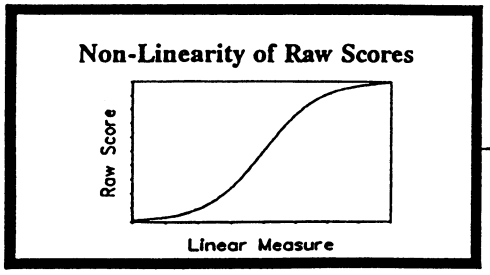
Edited by

John M. Linacre
University of Chicago

Based on
Rasch Measurement Transactions
Vol. 1:1 - Vol 4:1
Edited by Richard M. Smith
Vol. 4:2 - Vol 5:4
Edited by John M. Linacre

with the guidance of
Benjamin D. Wright

MESA Press
Chicago
1995



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 4 No. 3 | Autumn 1990 | ISSN 1051-0796

Objectivity: Specific and General

Objectivity is a cornerstone of all measurement. It requires that agent calibrations (e.g., item difficulties) be independent of the sample of objects (e.g., persons) used in the calibration procedure. Object measures must also be independent of the particular agents used to obtain them. Thurstone (1928) stated: "...the scale values of the statements of opinion must be as free as possible, and preferably entirely free, from the actual opinion of individuals or groups." As Wright & Stone (1979) remark, Loevinger (1947) offered a similar formulation as a criterion for absolute scaling, but it was Rasch (1961) who made what he called "specific objectivity" the central requirement of a new approach to measurement.

Specific Objectivity

Specific objectivity requires that *differences* between pairs of object measures or pairs of agent calibrations are sample independent. This means that two agents must be found to differ by the same (i.e., a statistically equivalent) number of measurement units no matter what sample of objects actually responds to the agents. Similarly, two objects must be found to differ by the same number of units no matter what sample of agents (from the universe of relevant agents) is used in the measurement procedure. In other words, the *relative* locations of pairs of objects and pairs of agents on the underlying continuum must be sample independent.

General Objectivity

General objectivity, essentially attained by measures in physics and chemistry (e.g., thermometers), requires that the *absolute* location of an object on, say, the Celsius scale, is sample independent. Temperature theory is well enough developed that routine manufacture of thermometers occurs without even checking the calibrations against objects with known values prior to shipping the instruments to customers. Such is our collective confidence in temperature theory. We know enough about liquid expansion coefficients, the gas laws, glass conductivity and fluid viscosity to construct a remarkably precise measurement with recourse only to theory. Measurement of the temperature of two objects results in not just sample independence for the difference between their temperatures, but sample independence for the point estimate of each object's temperature reading. It does not matter what thermometer we use, or how it was constructed, the Celsius value will be the same.

The difference between specific and general objectivity is seen not to be a consequence of the fundamental natures of the social and physical sciences, nor to be a necessary outcome of the method of making observations, but to be entirely a matter of the level of sophistication of the theory underlying the construction of the particular measurement instruments.

Jack Stenner
ComputerLand
2222 E. Highway 54
Durham NC 27713

Table of Contents

Assessment of Schizophrenia (W Fisher) . 113
Calendar of Events 118
CAT Fit Analysis (B Bergstrom) 112
Come to IOMW6 113
DIAL Lives On! (C Mardell-Czunowski) . 117
Directness and Measurement (T Rehfeldt) 117
Faulty Thinking (N Bezruczko) 114
Glossary of Misleading Terms (B Wright) . 116
History of Measurement (G Engelhard) . . 118
Mathematics of Rasch Model (P Alvarez) . 112
Note from the Chair (D Andrich) 115
Objectivity (J Stenner) 111
Scoring Patient Simulations (E Julian) . . . 116

RASCH MEASUREMENT TRANSACTIONS
PART 2

Edited by

John M. Linacre
University of Chicago

Based on
Rasch Measurement Transactions
Vol. 6:1 - Vol 8:4
Edited by John M. Linacre

with the guidance of
Benjamin D. Wright

MESA Press
Chicago
1996

Three Stages of Construct Definition

The development of construct definition follows a process that is articulated by its source of knowledge.

Stage 1) Instrument calibration based on personal knowledge, intuition and subjective analysis.
Pre-Galilean discussions of temperature measurement are interspersed with references to subjective "scales" of measurement anchored by terms like "as cold as when it snows" or "too hot to touch." A recent example is the attempt to measure "health risks of exposure to ionizing radiation." The observation (quantity of ionizing radiation) is converted into a measure (health risk) via calibrations based on the observer's value system. Objective measurement of constructs in their formative stages is difficult because theory is weak.

Stage 2) Data-based instrument calibrations.
17th Century temperature measurement employed data-based calibration. In Europe, two dozen "scales" competed for favor. Calibrations of thermometers were done on an instrument-by-instrument basis in the laboratory of the instrument maker. The particular readings of the thermometer, when exposed to states with known temperatures (e.g., human temperature), were used to calibrate each thermometer as it was manufactured. Measures from the same instrument maker were consistent and "specifically objective", i.e., two instruments from the same maker produced basically the same numbers. Measures from thermometers built by different instrument makers differed and there was no common frame of reference to permit a measure's reexpression in another metric.

A recent example of second stage construct definition is "mathematics achievement." Numerous instruments (tests) exist for measuring "mathematics ability", each with its own scale. Fifty years of factor-analytic research imply that all instruments measure something in common, but there is no shared framework that permits reexpressing one measure (e.g., NAEP) in terms of another (e.g., CAT). The confusion produced by multiple metrics contributes to the lack of consensus about what is, or should be, measured under the label of "mathematics ability".

Stage 3) Theory-based instrument calibration.
Thermometers made today are manufactured and shipped to customers without reference to data on the performance characteristics of the particular instrument. Instrument calibration is accomplished via theory-based equations and tables. Manufacturing proceeds with total reliance on theory. Theory enable. any measure to be reexpressed in the metric of another instrument maker (e.g., Celsius to Fahrenheit). Measures calibrated by theory are "generally objective." Any two observers given the same observation (volume displacement of mercury in a tube) will report back the same number as a measure.

The only behavioral science construct that approaches third stage development is "reading comprehension." This is because the Lexile Framework enables generally objective, theory-based, measurement of reading comprehension. Reading comprehension tests can be calibrated on the same metric, without reference to the performance of actual readers. The only reference required is the Lexile equation.

A. Jackson Stenner & Ivan Horabin
28 Stoneridge Circle, Durham NC 27705

MID-WEST OBJECTIVE MEASUREMENT SEMINAR

Friday, December 4, 1992, Chicago

"Discovery" in the Observation Model
Bonnie Roe, Museum of Science and Industry
Building "Architectural" Measures
William J. Boone, Indiana University, Michael Gorski, New York Institute of Technology
Which "Improvements" Alter Item Difficulty?
Betty Bergstrom, Computer Adaptive Techn.
"Unfolding" and the Dead Sea Scrolls
John M. Linacre, MESA Psychometric Lab.
Performance Profiles of the Functional Independence Measure (FIM™)
Carl V. Granger, SUNY at Buffalo
Computer-Assisted versus Written Tests
Mary Lunz & Greg Stone, Amer Soc Clin Path
Measuring Rehabilitation Client Satisfaction
Patricia Carter & William Fisher, Jr
Marianjoy Rehabilitation Hospital
Equating Exams with Few Examinees
Linjun Shen
National Board Osteopathic Medical Examiners
The Absolute Zero of Reading and Math
Ong Kim Lee, Malaysian Ministry of Education
Empirical versus Judgmental Item Banking
Bahrul Hayat, Indonesian Ministry of Education
Using Measures to Predict Amount of Care
Allen Heinemann, Rehabilitation Inst of Chicago

Using the Rasch model to validate a district-wide, curriculum-based mathematics assessment

This study describes a technique designed to help districts establish the instructional growth of their students using curriculum-based tests. Using Rasch methodology, the performances of 3rd and 6th grade students, pre-instruction (Spring) and post-instruction (Fall), were located on a calibrated variable, rather than reported as grade-equivalents, percentiles, or percent of outcomes mastered. Item and person fit patterns provided further group and individual information. These fit patterns were usefully categorized as item-related, person-related, or person-item-interaction characteristics.

Robert K. Hess (42.53)

Measuring values to apply the Golden Rule

This paper proposes a research program that would prepare the ground for a political morality based on the Golden Rule. This requires some way of discovering that "what I do unto others" is the same as 'what I would have done unto me." To discover this requires a measuring system that keeps things in proportion by showing what counts as "the same thing" for different people. This measuring system sets up analogies between people's values and what is valued. The measurement system is based on the specification that "my values are to one aspect of a situation what yours are to that or another aspect", and that proportions of this kind hold constant no matter what particular persons are addressed and no matter which aspects of the situation are involved.

William P. Fisher, Jr. (42.53)

While in New Orleans, You are Invited...

o **Mulate's**, the restaurant that made Cajun cuisine amous. Rasch folks and friends will dine on **Monday, April 4, 1994 at 7:00 p.m.**. Feel free to arrive early and gather in the bar before dinner. Mulate's has been ationally recognized for its celebration of Cajun ulture. Chef *Junior Savoy*, with 35 years of Cajun ooking expertise, has won awards from the American Culinary Federation. Mulate's is at 201 Julia Street 504/522-1492), right across Convention Center Boulevard from the Ernest N. Morial Convention Center. 80% of the distance (12 blocks) from the Sheraton to Mulate's is accessible by street car ($1.00 are). See you there!

William P. Fisher, Jr.

Construct Generalization

Construct Generalization is a method for investigating the fit of alternative construct theories to item difficulties obtained from tests purporting to measure the same or a similar construct. A good construct generalization study demonstrates how a number of different operationalizations of a construct (i.e., item calibrations) can be integrated and understood within a common theoretical perspective. During this process, the construct, the construct theories, and the test selection criteria are all improved.

For each construct of interest, we ask the simple question: "Is there a single equation, generated from a construct theory, that can account for the variation among item difficulties taken from a diverse set of instruments purportedly measuring this construct?"

The method of construct generalization consists of seven steps: (1) assemble a sample of relevant instruments, (2) estimate Rasch item difficulties separately for each instrument, (3) conceptualize alternative construct theories that explain variation in item difficulties and express these theories as specification equations, e.g., "reading difficulty is proportional to sentence length," (4) compute correlations separately for each instrument between the item difficulties as observed and those computed from the competing specification equations, (5) select a provisionally "best" construct theory and associated specification equation, (6) test the causal status of the variables in the specification equation by verifying that the specification equation predicts the observed calibrations for as yet untested items, (7) re-cycle steps 3-6 until satisfied with both the data-theory fit and the causal status of the variables in the specification equation.

A. Jackson Stenner
Computerland
1100 Perimeter Park West Suite 112
Morrisville NC 27560

From P-values and Raw Score Statistics to Logits

Conventional statistics and examinee raw scores may be all that remain of the data from a previous testing. These statistics provide the mean of the N examinee raw scores, their standard deviation and, for each item i, the rate of success of the sample of examinees, a "p-value", P_i, for $i = 1$ to L items.

In order to equate examinee performances on this earlier test with those on other tests, or to add these earlier items to an item bank, conversion from the raw score metric to a linear metric is needed. This can be achieved with a simple, usefully accurate technique, provided that it is reasonable to think of the examinees as randomly selected from a normal distribution.

Here's how to construct logit measures from raw score statistics:

1. Check the raw score statistics for consistency:

   Is the examinee mean raw score $\approx \sum_{i=1}^{L} P_i$?

Are there typographical errors?

Even if p-values were obtained from one sample and mean examinee score from another, they may still be close enough for this computation.

2. Compute a raw score-to-ability conversion factor, C_b:

$$C_b = \frac{1}{\sum_{i=1}^{L} P_i(1-P_i)}$$

3. Compute a logit examinee ability variance from SD, the examinee raw score standard deviation:

$$V_b = C_b^2 * SD^2$$

4. Obtain an item calibration expansion factor, X_f, to adjust item difficulties for examinee ability variance:

$$X_f = \sqrt{1 + \frac{V_b}{2.9}}$$

5. Compute a logit difficulty calibration for each item, d_i:

$$d_i = X_f * \log\left(\frac{1-P_i}{P_i}\right)$$

6. The standard error, SE_i, of calibration d_i for item i is:

$$SE_i \approx X_f * \sqrt{\frac{1}{N*P_i(1-P_i)}}$$

7. Compute an initial logit ability estimate b_r° corresponding to each raw score r from 1 to L-1:

$$b_r^\circ = X_f * \log\left(\frac{r}{L-r}\right)$$

8. Compute a final ability estimate b_r corresponding to each raw score r from 1 to L-1 by iterating the equation:

$$b_r \leftarrow b_r^\circ + \frac{r - \sum_{i=1}^{L} P_{ri}}{L * 0.25}$$

where

$$P_{ri} = \frac{1}{1 + e^{(d_i - b_r)}}$$

replacing b_r° by b_r and recomputing, until

$$\left| r - \sum_{i=1}^{L} P_{ri} \right| \le 0.25$$

9. For examinees with extreme scores of 0, compute b_0 using steps 7 and 8 with r = 0.25.

10. For examinees with extreme scores of L, compute b_L using steps 7 and 8 with r = L - 0.25.

11. The standard error, SE_r, of person ability measure b_r, corresponding to raw score r, including r = 0.25 and r = L - .25, is:

$$SE_r \approx \sqrt{\frac{1}{\sum_{i=1}^{L} P_{ri}(1-P_{ri})}}$$

This algorithm has been applied successfully to empirical data.

A. J. Stenner, B. D. Wright & J. M. Linacre

Specific Objectivity — Local and General

Georg Rasch used the term "specific objectivity" to describe that case essential to measurement in which "comparisons between individuals become independent of which particular instruments — tests or items or other stimuli — have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class — measuring the same thing — independent of which particular individuals, within a class considered, were instrumental for comparison." (1980 p. xx).

Local objectivity is the term used to designate the case in which **relative** measures are **empirically** discovered to be independent of which instrument is actually used to take the measures. Local objectivity is a consequence of a set of data fitting the Rasch model. When data fit, differences among person measures and among item calibrations are independent of one another, and hence of the sampling of items and persons. Fit means that any two items can be shown always to differ by a statistically equivalent amount no matter which sample of persons actually respond to the items. Similarly, any two persons can be shown always to differ by the a statistically equivalent amount, no matter which sample of items is used to implement the measurement procedure. Consequently, when data fit the Rasch model, then the relative locations of persons and items on the underlying continuum for a construct are independent of their sampling. Absolute measures can only be obtained indirectly by introducing some reference persons or reference items of specified absolute measure into the analysis.

Since local objectivity is empirically based, it can only be statistically confirmed by further sampling. Sampled results may imply a promising variable — promising because it spaces items into a useful substantive hierarchy of "meaningful moreness". Results may encourage, even confirm, a strong intention of how items "ought to" order. But the numerical specifics of the item difficulties are estimates form this sample of persons. They may turn out, on further sampling, to be statistically reproducible, but their quantities cannot be deduced other than by reference to empirical data.

An ideal, approximated by instruments in physics and chemistry, is that **absolute** calibration of an instrument, such as a thermometer, does not require the services of an object or another instrument. Further the location of an object on, say, the Celsius scale, is not instrument dependent, i.e., any "thermometer" will do. Neither does that location depend on measuring any other object. Temperature theory is well enough developed that thermometers can be constructed and calibrated without reference to any object or any other thermometer. Measurement of the temperature of two objects results in not just instrument independence for the difference between their temperatures, but also instrument independence for the amount of each object's temperature measure.

Absolute measures are obtained directly when a specification equation, which implements a theory, calibrates instruments with useful and reproducible precision. The instrument calibrations are then based, not on the measurement of any objects, but on the design and efficiency of the specification equation.

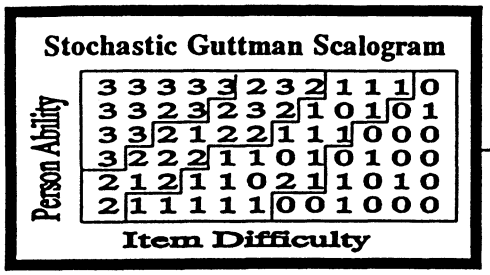
The term *general objectivity* is reserved for the case in which absolute measures (i.e., amounts) are independent of which instrument (within a class considered) is employed, and no other object is required. By "absolute" we mean the measure "is not dependent on, or without reference to, anything else; not relative" (Webster 1972).

The Table compares **local** and **general** objectivity. "The difference between *local* and *general* objectivity is seen not to be a consequence of the fundamental natures of the social and physical sciences, nor to be a necessary outcome of the method of making observation, but to be entirely a matter of the level of theory underlying the construction of the particular measurement instruments." (Stenner, 1990, *RMT* 4:3 p. 111).

A. Jackson Stenner
Metametrics Inc.
1100 Perimeter Park West
Morrisville NC 27560

Anatomy of Objectivity		
Aspect	**Local objectivity**	**General objectivity**
Basis	empirical	theoretical
Data	sampled, exploratory, effects unknown	constructed, specified, effects known
Philosophy		
Intention	exploration	measurement
Construct Definition	incomplete, data-discovered	complete, theory-specified
Observations	quantified by data-based inference	quantified by theory-based specification
Origin, Unit, Precision	sample-estimated, varies	theory-specified, fixed
Item Calibration		
Calibrations	sample-estimated differences	theory-specified amounts
Misfit Diagnosis	depends on person and item sampling	construct consistency
Person Measurement		
Measures	sample-estimated differences	theory-calculated amounts
Misfit Diagnosis	item-by-person confounded	person-specific
Meaning		
Criterion	implied by sampled items	defined by theory
Norms	sample and test specific	general to the scale
Meta-analyses	aggregates indices of relative effects	aggregates amounts in measured units
Manufacturing		
Test Costs	person/item sampling, many iterations, expert supervision & evaluation	person and item targeting, routine review
Equating & Banking	sampled from common persons or items	pre-specified by common theory
Quality Control	expert evaluation, person-by-item confounding	pre-designed routine

A. Jackson Stenner



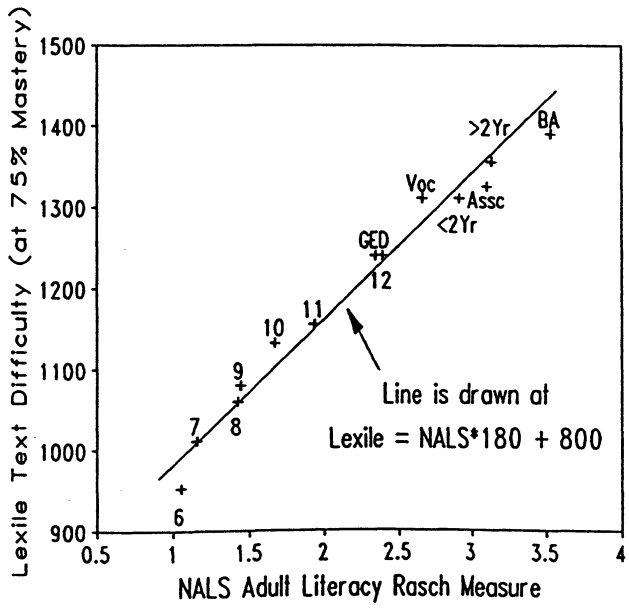
RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 8 No. 4 | Winter 1995 | ISSN 1051-0796

Reading in America: Stenner's Lexiles Confirmed!

Progress is signalled by the ability to predict outcomes. Data from the 1992 National Adult Literacy Survey (NALS) confirm Jack Stenner's (1987) *Lexile* theory in the same way that quark detection confirms particle physics. When the lexile values necessary for 75% mastery of the reading material used at each grade level are plotted against the mean NALS Rasch measures of adult literacy corresponding to each year of final education, a startlingly simple and decisive confirmation of the *Lexile* theory appears. The relationship between the theoretical lexiles and the empirical Rasch measures is a **straight line!** It follows that the *Lexile* construct can organize and predict a fundamental feature of our complex world, the ability to understand and use written language. The educational, commercial and political implications are profound and extensive.



Lexile theory posits that the reading difficulty of ordinary text is dominated by word frequency and sentence length. Stenner's *Lexile Scanner* reads text segments, looks up word frequencies and counts sentence lengths. It then predicts the reading difficulty of the text in *lexile units* computed from the mean log word infrequency and the log mean sentence length. Stenner has lexiled reading materials from nursery to graduate school and beyond. This includes the particular materials used at each level of education from grade school through college.

The NALS data come from 24,944 US adults answering selections from 169 English literacy items. Dick Vanezky, Mike Linacre and Ben Wright have estimated the mean Rasch NALS literacy levels for adults at each year of final education from grade school through college. The NALS items were designed and administered with no reference to *Lexile* theory.

In the plot, mean literacy levels increase in equal steps with education levels. Thus, on the average, we are

Table of Contents

AERA, IOMW8 Sessions and Abstracts . . 389
Assigning Item Weights (Linacre, Wright) 403
Bruce Choppin (J Linacre) 394
Clients vs. Therapists (Gerardi, Eckberg) . 399
Dimensional Analysis (M Moulton) 398
Improve Presentations (BAR) 388
Medical Diagnosis (Cristante, Robusto) . . 395
Polytomous PROX (J Linacre) 400
Problem Drinking (M Cornel, R Knibbe) . 402
Reading and Lexiles (B Wright) 387
Recalibration Stability (Lunz, Bergstrom) . 396
Truth from Fiction (W Fisher) 401
WRAT3 Item Map (M Stone) 403

only as literate as the reading materials at our last year of education! Our literacy level does not change with reading activity after formal schooling ends, but our literacy does decline from around age 40. Other NALS analyses show a similar linearity between mean literacy, education level and log income! Thus, on the average, we are **only as rich as our literacy and education allow!**

Stenner's studies show that the literacy level needed to read newspapers is above 1200 lexiles. On the plot 1200 corresponds to finishing high school. Thus, **for a democracy based on informed voters to survive, we must see to it that all youth finish high school.** When we leave school prematurely, we condemn ourselves to an unnecessarily impoverished life. **Staying through high school has got to be a national priority!**

Benjamin D. Wright

Carol Myford receives ETS Research Scientist Award

ETS President Nancy S. Cole awarded ETS Research Scientist Awards on December 19, 1994, to Carol Myford and Neal Thomas. Nancy Cole stated that "I have great pleasure in announcing that Carol Myford has been selected to receive the ETS Scientist Award. This award recognizes ETS researchers who have distinguished themselves through outstanding contributions to their fields of specialization and to the work of ETS... Carol's work with the NAEP Arts Education Consensus Project Team has identified ETS with a nationally recognized effort using new modes of assessment."

Tony Cline, executive director of the Division of Applied Measurement Research nominated Myford for the award: "Carol was able - with extraordinary skill, expert knowledge and diplomacy - to get a group of 32 arts educators and performing artists to produce sets of assessments and exercise specifications for NAEP for grades 4, 8, and 12 in dance, music, theater (including film, television, and video) and visual arts (including design, architecture and media arts. She was the only ETS staff member involved in the effort. She is now accepted as an expert in the area of using rating scales to evaluate live performances." Myford commented "For me, the NAEP Project was the most difficult writing assignment I have ever undertaken, but it was also the most rewarding collaboration I've ever been involved in. The Project was truly a team effort."

Quotations from ETS Access 5(15) 12/15/95

Let's Improve our Presentations!

Here are 7 guidelines for improving the quality of paper presentations suggested by Herschel Shanks in *Biblical Archaeology Review (1993, 19:2 p.50).*

1. **Prepare your paper for oral presentation.** Most papers start out from a text meant to be read rather than spoken. We speak differently than we write. We hear differently than we read. Say it out loud. Does it sounds natural?

2. **Speak, don't read, to your audience.** This may take practice, but it's worth it. People in the audience have often come thousands of miles to hear you.

3. **Be informal.** It is really possible to be informal and scholarly at the same time.

4. **Time yourself.** When the chair tells speakers they're out of time, it's embarrassing. Worse, speakers are prevented from making their most important points. There is an easy antidote: time yourself beforehand.

5. **If you must summarize previous work,** be brief and to the point. Your audience assumes you've done your homework. We've come to hear you tell us what is new.

6. **In your opening sentence, tell your audience what your conclusion is going to be.** If we know where you're going, we will follow you better and be less likely to doze.

7. **Edit your presentation beforehand.** After you've finished preparing your talk, read it over carefully and ask yourself if there are any topics that can be eliminated without damaging the flow of your argument. To understand your main points, do we need to know the sample demographics? The estimation equations? Your previous findings? If not, eliminate them. Then present your material aloud to yourself, asking yourself if your are presenting your material logically, clearly and precisely. Less complexity gives more impact.

In what matter soever there is place for addition and subtraction, there also is place for reason; and where these have no place, there reason has nothing at all to do. *Thomas Hobbes (1588-1679)*

Of Reason and Science. Leviathan, Part I, Chap. V

Courtesy of Thomas K. Rehfeldt