# Building Programs, Evaluations, and Tools to Promote Human Flourishing in LMICs:

# Unasked Questions and Welcomed Answers

**TWCF Webinar 3:**

**December 10, 2020**

Richard M. Lerner, Marc H. Bornstein, and
Elizabeth M. Dowling

# The Goals of Character Virtue Development Programs in LMICs

- **To enhance program participants' character virtues and other indicators of positive development – such as PYD, SEL, or thriving or flourishing.**

- From preceding TWCF webinars, many ideas that can be useful for designing and evaluating effective programs have been presented.

- Useful tools exist for indexing specific instances of flourishing, character virtue development, and related constructs of relevance in specific communities in specific countries.

# The Goal of This Webinar

- Prior webinars have provided useful answers to commonly asked questions about program design, evaluation, and the validity and reliability of measurement. However, some important issues still need to be discussed.

- Accordingly, in this final webinar we take the bold step of providing answers to some important ***but unasked*** questions.
- The questions and answers concern specific issues that revolve around validity and reliability in:

  - Randomized Control Trials (RCTs)
  - Measurement
  - Missing data
  - Using group averages to represent specific people in specific programs

# How Deep Can We Go?

- Of course, in the time we have available for this webinar, we can only touch the surface of these questions and answers.
- Therefore, we invite you to work through the TWCF to ask for additional information, clarification, consultation, or support.
- Let's get started by first defining reliability and validity.
- We will then ask a question that Marc Bornstein answered in his webinar, but we think needs to be answered again – about the validity of traditional approaches to conducting RCTs.

# The Many Meanings of Reliability

- **Reliability** is consistency in observations. Reliability occurs when two observations yield the same result (e.g., a measure of the character virtue of gratitude yields the same scores on two different occasions).

  - *In experimentation*, reliability means that the results of an experiment can be replicated.

  - *In survey research*, reliability means that a correlations between two constructs (e.g., between purpose and generosity) is found in different samples.

  - *In measurement*, reliability means that items (e.g., indexing humility) are interrelated ("internally" consistent), that scores for the items are the same across different times of measurement, or that two independent raters provide the same scores.

  - *Also in measurement*, reliability is the consistency in scores from two measures that are designed to be highly similar.

# The Many Meanings of Validity

- **Validity** occurs when the observation represents what it is intended to represent. Validity exists when there is evidence that the construct (or "latent variable") that a researcher intends to measure is actually in evidence (e.g., there is evidence that a measure of the character virtue of gratitude actually measures gratitude).

  - *In experimentation*, validity means that the outcome of the experiment is due to, and ONLY due to, the manipulation or treatment, and not to any other variable or event. **This situation indicates internal validity**.

  - In *survey research*, validity means that a measure (e.g., of generosity) is correlated with or predicts something that occurs when the construct (generosity in this case) actually exists (e.g., the magnitude of anonymous annual charitable giving).

  - *In measurement*, validity points to true indicators of the construct (or latent variable) a researcher is trying to measure.

  - *Also in measurement*, validity exists when there is a strong relation between two measures of a construct that are designed to be maximally different.

# Question 1

- Are RCTs really the gold standard in program evaluation?

- If not, what alternatives are available?

# Alternatives to the Traditional (2-group) RCT

- **The Solomon 4-Group Design**

- **Econometric Alternatives to RCTs:**
  - Counterfactual comparative structural equation modeling using propensity score matching
  - Regression discontinuity designs
  - Instrumental variable analysis

# The Solomon 4-Group Design

Solomon, R.L., & Lessac, M.S. (1968). A control group design for experimental studies of developmental processes. *Psychological Bulletin, 70,* 1545–1550.

# Randomized Control Trials (RCTs): The Gold Standard?

- As implemented through the first decades of the 21st century, many RCTs were "Fools' Gold."
- The entire purpose of an RCT is to assure internal validity.
- RCTs are internally valid *if and only if* pre-existing attributes of participants (termed selection effects or "endogeneity") are excluded as possible sources for variation in the outcomes of treatment or intervention.
- The "traditional" design of the RCT has only two groups – a treatment (or intervention) group and a control group.
- **However, not one, but three, control groups are needed to assure internal validity.**

## The Solomon and Lessac (1968) 4-Group Design: Overview of the Groups Needed for Internal Validity

| | Experimental or Treatment Group | Control Group 1 | Control Group 2 | Control Group 3 |
|---|---|---|---|---|
| Pretest | ✓ | ✓ | | |
| Treatment | ✓ | | ✓ | |
| Posttest | ✓ | ✓ | ✓ | ✓ |

# Explaining the Logic of the *4-Group Design*

| | Intervention Groups | Control Groups |
|---|---|---|
| Condition | 1 | |
| Pretest | ✓ | |
| Intervention | ✓ | |
| Posttest | ✓ | |

Pretest determines level of participants' attributes before intervention

Pretest-Posttest comparison determines effects of intervention … improvement

# Explaining the Logic of the 4-Group Design

| | Intervention Groups | | Control Groups |
|---|---|---|---|
| Condition | 1 | 2 | |
| Pretest | ✓ | -- | |
| Intervention | ✓ | ✓ | |
| Posttest | ✓ | ✓ | |

Now the Pre- and Posttests are substantively related to the intervention and so the Pretest gives relevant experience, and the Posttest=Pretest, so perhaps the Pretest is the "active ingredient", not the intervention …

Group 1 v. 2 comparison shows/eliminates effects of Pretesting vs Intervention

# Explaining the Logic of the *4-Group Design*

| | Intervention Groups | | Control Groups | |
|---|---|---|---|---|
| Condition | 1 | 2 | 3 | Group 3 with no intervention also tests the Pretest priming |
| Pretest | ✓ | | ✓ | |
| Intervention | ✓ | ✓ | | Comparing Groups 1 v. 3 on the Pretest-Posttest difference reveals the unique or non-effect of the intervention |
| Posttest | ✓ | ✓ | ✓ | |

# Explaining the Logic of the *4-Group Design*

| | Intervention Groups | | Control Groups | |
|---|---|---|---|---|
| Condition | 1 | 2 | 3 | 4 |
| Pretest | ✓ | | ✓ | |
| Intervention | ✓ | ✓ | | |
| Posttest | ✓ | ✓ | ✓ | ✓ |

Between the Pretest and Posttest in Group 1, participants grow and change… Groups 1 v. 4 compare intervention with development; "Hawthorne Effect"; and diffusion or contamination

# Explaining the Logic of the *4-Group Design*

| | Intervention Groups | | Control Groups | |
|---|---|---|---|---|
| Condition | 1 | 2 | 3 | 4 |
| Pretest | ✓ | | ✓ | |
| Intervention | ✓ | ✓ | | |
| Posttest | ✓ | ✓ | ✓ | ✓ |

# Options for Comparison or Control Groups (Bornstein, 2020)

- The ***No Treatment*** (or ***Business-as-Usual***) **condition**: Participants are offered no participant-related support as part of their study participation.

- The ***Nominal Support*** **condition**: Participants are offered a modest level of assistance related to, but not part of, the actual support or intervention design.

- The ***Treatment*** **(or *Intervention*) Component condition**: Participants are offered an element or lower dosage of the intervention.

- The ***Attention Placebo*** **condition**: Participants are offered contact with the intervention staff at the same intensity as participants in the actual support and intervention minus the  focus of the intervention.

- The ***Waitlist Comparison*** **condition**. Participants receive the treatment condition but wait (with no treatment) to receive it until the end of the assessment of the original treatment (intervention) group. A waitlist control group can be used to assess the replicability of the treatment. If so, then the initial treatment group and the waitlist control group need to be treated identically.

# Econometric Alternatives to RCTs: 1. Counterfactual comparative structural equation modeling using propensity score matching

- There may be practical or ethical constraints in assigning people to participate in specific programs (e.g., involving character virtue education) and, of course, people cannot be randomly assigned to race, gender, poverty, or religious groups.

- However, people who are in a program also have specific race, gender, poverty, religious, etc. attributes. These attributes exist also among people who are not in the program.

- This second group of people are called *counterfactuals*.

- A propensity score is the probability of a person being assigned to a group based on all the variables that encompass a group, either the program group or the counterfactual group (these variables are termed *covariates*).

- Covariates can be used to create a counterfactual sample matched to the program sample through propensity score matching, thereby reducing *selection bias* (*endogeneity*).

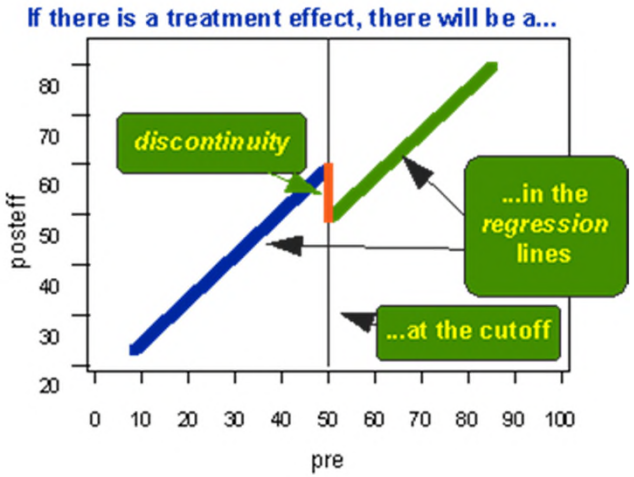
- Statistical procedures (e.g., *structural equation modeling*, or *SEM*) can be used to model the causal effects of program participation (e.g., on character virtue development).

# Econometric Alternatives to RCTs: 2. Regression Discontinuity Designs

- A pretest cutoff score is used to assign participants to either the program or comparison group.

| | | | |
|---|---|---|---|
| C | O | X | O |
| C | O | | O |

- The assumption is that in the absence of the program the pre-post relation would be equivalent for the two groups (William Trochim, 2006).



# Econometric Alternatives to RCTs: 3. Instrumental Variable (IV) Analysis

- An instrument is a variable that does not itself belong in the explanatory equation and is correlated with the endogenous explanatory variables, conditional on the other covariates.

- In attempting to estimate the causal effect of some variable *x* on another *y*, an instrument is a third variable *z* which affects *y* only through its effect on *x*.

- For example, to estimate the causal effect of ("*x*") smoking on ("*y*") general health, one may use the tax rate on tobacco products ("*z*") as an instrument for smoking in a causal analysis: If tobacco taxes (*z*) affect health (*y*) *only* because they affect smoking (*x*) (holding other variables in the model fixed), the correlation between tobacco taxes and health is evidence that smoking causes changes in health.

# Question 2

- These alternatives show that you can create compelling evidence for the effectiveness of a program in several ways different from conventionally designed (2-group) RCTs.
- Designing evaluations is one thing. Using measures that reflect desired outcomes is quite another.

- **Given my limited resources of staff, time, and money, must I develop measures that have meaning, validity, and reliability in my specific setting and with the specific individuals in my program?**

- **Why can't I just use measures that have been used before, especially if they have been shown to be reliable?**

# Reliability and Validity in Measurement

**Reliability and validity are attributes of specific participant's behaviors** ***and not of the words on a survey*** (Noel Card, 2017).

- People are reliable or not reliable. People's actions are or are not valid indications of their true thinking, feeling, or behavior. **THEREFORE, you must establish reliability and validity for the specific groups with whom you are working!**
- A measurement tool may or may not provide an observation that is a true (valid) indication of an individual's functioning at one point in time or of the person's development.
- Measures must be change sensitive. Measures must also be relevant to specific developmental levels and to specific cultural contexts of participants.
- Most important, measures must be indictors of reliable **AND** valid attributes of an individual.
- Unfortunately, many researchers trade validity for higher reliability.

# Trade-Offs Between Reliability and Validity

**(Card, 2017; Clifton, 2019)**

- The purpose of measurement is to obtain a true index of a facet of an individual's behavior and/or development.
- However, all measurement involves variation from three sources:
  **1. True assessment of a specific observable behavior or of a "latent" construct,**
  **2. random error, and**
  **3. systematic error.**
- **Random error** can occur when a person misreads a word, is distracted by a fly landing on his head, or there is an electrical outage when taking an online survey.
- **Systematic** error occurs when the method of administering a measure biases participants' responses to the measure.

***Importantly, systematic error tends to increase reliability of responding but decreases validity of responding.***

# Sources of Systematic Error: Slide 1

- **Following Clifton (2019), some sources of systematic error – and of the *tradeoff between reliability and validity* – include:**

1. ***Item language problems***: Similar words or grammatical structures introduce systematic error associated with those words or structures. For example, if a 10-item scale has 5 items that begin with "I feel that ..." but the remaining 5 items have five quite different wording (e.g., I believe…, I think, … My view is…, etc.). The "I feel that ..."items are likely to correlate highly with each other and increase reliability—all driven by systematic error associated with language similarity and *not* measurement of the true construct;

2. ***Acquiescence bias***: Acquiescence bias (agreement bias) is the tendency for respondents to agree with items regardless of what the item is about;

# Sources of Systemic Error: Slide 2

3. ***Fixed-order bias:*** Administering items in a fixed order is problematic because each item serves as a prime for the next. Although sequence effects vary in magnitude, every item administered in a sequence inevitably influences those that follow in a nonrandom way. The result is increased communality among items, increased reliability, and weakened validity; and

4. ***Response bias:*** The results of demand characteristics as shown in consistent responding, deviant responding, careless responding, agreement bias, affect bias, and social desirability.

*5.* ***Item difficulty:*** Variance among items may be an artifact of specific levels of difficulty. For example, among items involving understanding moral reasoning principles and understanding mathematical concepts, there may be two groups of highly correlating items: hard items and easy items. Analyses may capture this variance in item difficulty and result in spurious dimensions, which are termed difficulty factors.

# How Systematic Error Can Affect Reliability Through Use of Cronbach's Alpha Coefficient

- Coefficient alpha is a standard measure of the internal reliability of a scale (that is, the covariation among items in a scale).

- Imagine a 9-item scale involving 5 items concerning gender and 4 items concerning hair color. If interitem correlations averaged .95 within the 5- and 4-item sets and .00 between sets, the average correlation across all 9 items would be .42, alpha would be *very good* at .87, and yet the scale as a whole would measure nothing.

- Therefore, incremental change in alpha indicates that the proportion of covariance among items is increasing *without providing insight into why*.

# Question 3

- So now we understand more about how to reduce systematic error and obtain estimates of reliability that increase true variance about the construct (the latent variable) we want to measure.

- **But, if validity is the goal – that is, if we want a true measure of the construct in which we are interested – what measure of validity should I use?**

# Validity

- Validity involves the correct identification of the latent variable (e.g., the character virtue of generosity, purpose, or humility), which is a conceptual (theory-based) determination of meaning.

- *That is, validity is subjectively determined*. No numerical indicators or "rules of thumb" are applicable to all instances of validity.

- Distinguishing high versus low validity is meaningless because validity is non-numerical (Anastasi & Urbina, 1997).

- In other words, validity is "in the eye of the beholder," or in the theoretical explanation of why a specific form of validity is useful.

# Types of Validity: Slide 1 of 2

- ***Content validity****:* The degree to which items denote the right construct, the entire construct, and nothing else. *Content validity* is what should determine scale labels.

- ***Predictive validity*** *(or* ***criterion-related validity****)*: The degree to which scale scores occupy the right spot in the nomological net or have the right (i.e., theoretically expected) empirical associations.

- ***Divergent and convergent validity:*** Scores are unrelated to (have no correlation with) variables presumed to be unrelated (*divergent/discriminant validity*) BUT are significantly correlated with variables presumed to be related (*convergent validity*).

# Types of Validity: Slide 2 of 2

- ***Incremental validity*:** The measure accounts for new variance in variables presumed to be related to the measure (e.g., a measure of forgiveness adds to the variance associated with other measures of character virtues).

- ***Concurrent validity*:** A measure is highly correlated with variables presumed to reflect the same latent phenomenon.

- ***Expert rater validity*:** Experts agree that a measure (or items) are indicators of a specific construct (latent variable).

- ***Factorial validity:*** The measure has a theoretically specified configuration of factors (or components).

# Question 4

- We now understand how to avoid reliability-validity trade-offs and different ways to establish validity.
- **What else do I need to know regarding how to measure the attributes of the individuals in my program?**
- For instance, is there anything I should know about how I ask people to respond to the items in my surveys? For example, is it okay to use a 5-point Likert response scale or is it always better to use a 7-point Likert response scale?

# Response Scale Options: TOWARDS THE END OF LIKERT SCALES

- On a Likert scale, respondents specify their level of agreement or disagreement on a symmetric agree-disagree scale for a series of statements.

- We advise to **stop using Likert scales**.

- The reason requires a brief excursion into the nature of measurement.

# Types of Measurement: Slide 1

- **Nominal:** Measurement by naming the categories or groups into which observations fall, for example, male-female, children-parents, children-adolescents, conservatives-liberals, etc.

- **Ordinal:** Measurement by putting observations into an order in relation to a specific dimension, for example, largest to smallest, tallest to shortest, most true to least true, or highest agreement to lowest agreement. *The differences among elements in this ordering need not be the same and, in many cases, is not specified or known. Likert scales are ordinal scales*.

# Types of Measurement: Slide 2

- **Interval:** Measurement using equal spacing for the ordering of observations, for example, from no difference in height to one, two, three, four, etc. inches or cm of difference in height; from no difference in weight to one pound, two pounds, three pounds, or kg, etc. differences in weight; or from zero percentage of time (0%) a specific observation is found to exist to, for instance, 31%, 63%, 79%, 84%, to 100% of the time an observation exists (e.g., for the item "I pray at night before going to sleep," a true interval response scale might be "0% = Never" to "100% = Always."

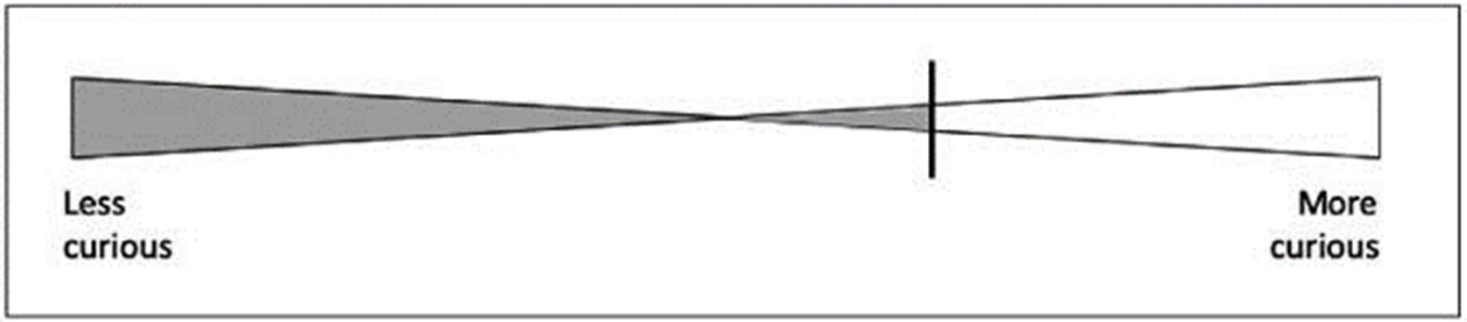
- **Ratio:** An interval measurement with a true zero, where zero indicates the absence of the quantity being measured (e.g., temperature). The ratio scale is the most informative scale: It integrates in one measurement the three earlier scales.

# Using an Interval Response Scale

- **Response Scale Coarseness:** This instance of measurement error will occur when measuring a character virtue, attribute of flourishing, or any other construct that is continuous in nature (i.e., that can vary to any degree from zero, or completely absent, to 100% present) with an ordinal scale of measurement, such as a Likert scale.
- Coarseness involves collapsing differences in true scores into the same category.
- This measurement problem is reduced as the number of answer categories increases and is avoided completely by using a truly continuous equal interval scale, such as the visual analog scale (Rioux & Little, 2020).

# The Visual Analog Scale

I am curious about science. (Less Curious= 0; More Curious = 100).



# Strengths of True Interval Scales

- The visual analog scale allows researchers to capture continuous constructs appropriately and respondents to freely specify their level of agreement with an item, instead of being constrained to predetermined categories.
- Data collected in studies using a visual analog scale are more precise and provide more information than data collected using the Likert scale.
- Moreover, the visual analog scale is valid, reliable, efficient to administer, and easy to use (and intuitive) for participants.
- Although the visual analog scale offers many advantages, implementation requires a computerized questionnaire.

# Question 5

- **We now understand how to enhance the reliability and validity of our measures. Is there something besides reliability and validity to know about measurement?**

# Yes!

- You need to also know if a measure that is valid for specific people, contexts, or times is valid for other people, contexts, and times.
- This is **measurement invariance**!
- Measurement invariance is a type of ***external validity*** or, simply, of generalization.
- ***IF*** a measure for one group of people (e.g., boys) works equally well for another group (girls), or ***IF*** a measure in one country (e.g., Rwanda) works equally well in another country (e.g., El Salvador), or ***IF*** a measure (e.g., of a character virtue) that works at the beginning of a program works equally well at the end of a program, ***THEN*** you have established measurement invariance.

# The Concept of Measurement Invariance [1]

- **Invariance means that a measure performs in the same way across:**
  - People (e.g., boys and girls, individuals in treatment vs. comparison groups),
  - Contexts (e.g., rural vs. urban, Country 1 vs. Country 2), and
  - Time (e.g., pre- vs. post-intervention, waves in a longitudinal study).

- **Other terms used to represent the concept of measurement invariance:**
  - Measurement equivalence
  - Factorial equivalence
  - Factorial invariance
  - In Item-Response Theory [or IRT] the absence of differential item functioning [DIF]

- Documenting measurement invariance, that is, that the measure has the same meaning, structure, and properties for different people, contexts, and time, requires use of factor analysis methods.

**[1] Based on Card (2016, 2017)**

# Summary of the Four "Levels" of Measurement Invariance

- *Equal form*: The number of factors and the pattern of factor-indicator (or factor-item) relations are identical across groups [**Configural Invariance**]

- *Equal loadings*: Factor loadings are equal across groups [**Metric or Weak Invariance**]

- *Equal intercepts*: When observed scores are regressed on each factor, the intercepts are equal across groups [**Scalar or Strong Invariance**]

- *Equal residual variance*: The residual variances of the observed scores that are not accounted for by the factors are equal across groups [**Strict Invariance**]

# Question 6

- We now understand how to measure character virtues with reliability, validity, and measurement invariance. But in evaluating my program, participants sometimes don't answer all the questions I ask them or, worse, some participants miss entire occasions of measurement. Simply, I have incomplete data – missing data – and I don't know how to determine if my program is effective if I have missing data.

- **What can I do about this?**

# Missingness

- This question needs to be answered in two parts.
- First, we need to review different types of missingness.
- Second, we need to discuss what can be done when there is missingness.

# Four Categories of Missingness

- **Structurally missing data** are data that are missing for a logical reason, data that are missing because they should not exist (e.g., if a study assesses children ages 5 to 7, there should be no data about 4- or 8-year-olds or if a response scale varies from 1 to 7, there should be no "8"s in the data set).

- **Missing Completely at Random** (**MCAR**). Whether a participant has missing data is completely unrelated to any other information (variables) within the data set (e.g., in a "*planned missingness design,*" some participants are randomly assigned to not being tested at Wave 1 or to having to respond to only ½ of the overall measures used in a study at a specific wave of testing).

# Categories of Missingness cont.

- **Missing at Random (MAR).** Missing data that can be predicted from other information (variables) in the data set. For instance, in a study of physical endurance across multiple trials (e.g., in research about athletic training), some individuals with lower lung capacity will be more likely to drop out of the study than individuals with higher lung capacity.

- **Missing Not At Random (Nonignorable Missingness**). Missing data because people in a sample opted to not participate because of specific attributes (e.g., in a study of the personality and adjustment of people who endorse racist ideology or attitudes of white superiority, people who believe that the study would reveal unfavorable attributes about them, might not participate).

# What Can Be Done About Missing Data? Slide 1 of 2

- Todd Little has famously observed that the best way to address problems of missing data is to not have any! However, missing data are ubiquitous in developmental research, especially in research or evaluation involving multiple times of testing.

- **List-wise deletion** (dropping from analysis any person who has any missing data) remains the most frequent missing data treatment. However, this method is rarely the best one to use. Deletion methods are associated with loss of statistical power and increases in sample bias.

# What Can be Done about Missing Data: Slide 2 of 2

- The recommended approaches are multiple imputation (MI) and full information maximum likelihood (FIML).
  - MI is a two-step approach in which missing values are first estimated to create imputed data sets, and analyses are then run on the imputed datasets.
  - FIML deals with the missing data by estimating parameters and standard errors in a single step.
- Both these missing data treatments have advantages and limitations, and researchers should consider many factors when choosing a missing data treatment (e.g., variable distributions, item aggregation, study design, non-response pattern, nonresponse rate, number of variables, number of observations, and type of analysis).

# Question 7

- We have learned a lot about how to improve measurement of variables, for instance, the specific character virtues I want to see developed in a participant in my program. However, I developed my program because I wanted to enhance the lives of each and every person in my program.
- I am told by program evaluators that my program will be a success if I raise the *average scores* of all the people in my program. That is, I am told to focus on average changes in scores for variables. I am therefore documenting changes in the scores of variables, and not of any specific individual.
- If I am really interested in enhancing the flourishing of each individual, then why am I studying variables and groups and not documenting if and how I am changing each specific individual in my program?
- How can I know if my program is helping specific people if all that I prove about my program pertains to a group average?
- **Do group averages = the specific scores of specific people in my program?**

# The End of Averages and the Rise of Specificity

- NO! Group averages do not necessarily = the specific scores of specific people in a program
- Changes in the average score for a variable for the group of participants in a program (e.g., the character virtue of generosity) do not tell you anything about the changes of any specific individual whose scores are part of the average.
- To illustrate, let's consider the apocryphal story of Napoleon's bootmaker:
- Napoleon wants to invade Russia. His army will need boots. He hires a bootmaker. The bootmaker correctly thinks he needs to know the size of Frenchmen's feet to order the leather. He measures a bunch of soldiers, gets the average, multiplies the average by the size of Napoleon's army, gets the right amount of leather, and makes all the boots the average size. Of course, no boots fit any individual soldier.

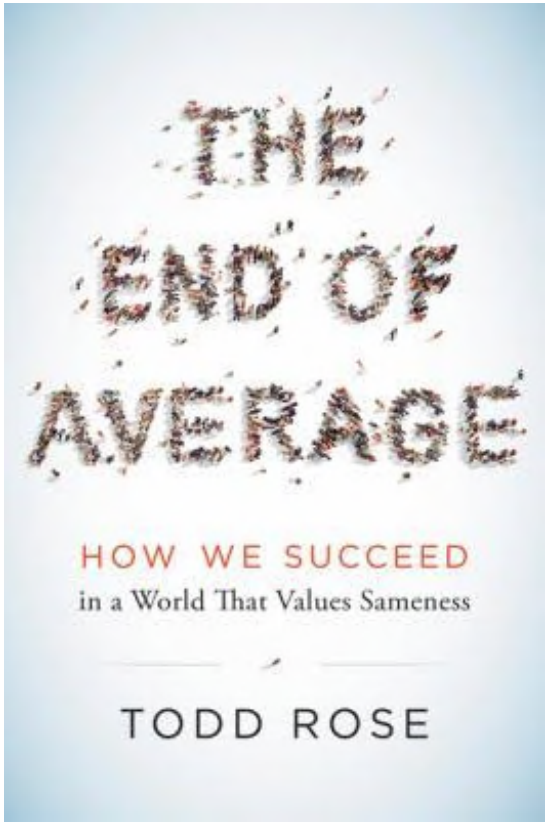
# Why We Use and Why We Abuse the Use of Averages

- Averages and standard deviations (which are variations around an average, or mean, score) are statistical tools that are useful to depict attributes of populations.
- These tools are founded in mathematical theorems, the **Ergodic Theorems**.
- These theorems enable averages and standard deviations for samples from a population (e.g., the participants in a character education program) to be computed.
- However, these statistics can only be computed if two conditions exist:
  - ***Homogeneity***: Each member of the sample must be the same (you cannot compute an average if some members of the sample are apples and others are oranges).
  - ***Stationarity***: Each sample member's variation around the mean does not change across time and place.
- If these two conditions are met, ergodicity exists, and then means and standard deviations may be computed.
- An interesting implication of ergodicity: It is assumed that the average for the group applies equally to all members of the sample, that is:
  - Each individual's score = the average score + random variation (error).

# Why we Use and Why we Abuse the Use of Averages

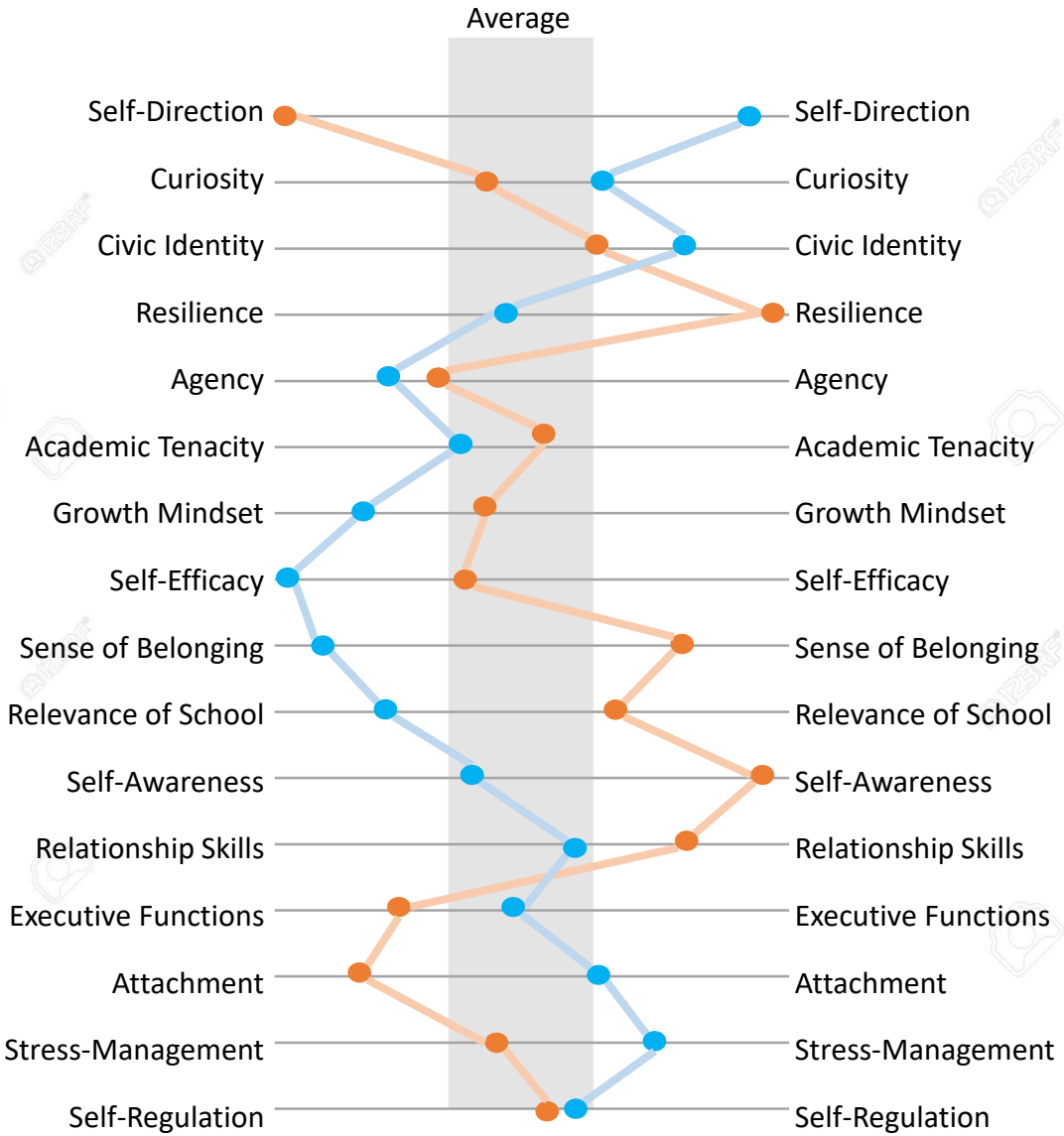
- Most developmental scientists would admit that, in reality, the people in the samples they study are not all the same.
- Most developmental scientists would admit that differences between the group average and the scores of sample members are not just error.
- Indeed, most developmental scientists would agree that **human development is non-ergodic**!
- However, many would still claim that averages are "good enough approximations" of individuals and, as such, there is no need to focus on the specifics of the individuals. They would say that for most of what we want to know about human development, focusing on average scores for variables is sufficient.

**IS THIS LINE OF ARGUMENT TRUE?**


THE
END OF
AVERAGE
HOW WE SUCCEED
in a World That Values Sameness
TODD ROSE

# Three Concepts that Go Beyond an Exclusive Focus on Average

1. ***Jaggedness***: At any point in time each person has his or her ***specific*** and potentially unique constellation of attributes (e.g., academic, moral, civic, social, and leadership).


Average
Self-Direction
Curiosity
Civic Identity
Resilience
Agency
Academic Tenacity
Growth Mindset
Self-Efficacy
Sense of Belonging
Relevance of School
Self-Awareness
Relationship Skills
Executive Functions
Attachment
Stress-Management
Self-Regulation
Self-Direction
Curiosity
Civic Identity
Resilience
Agency
Academic Tenacity
Growth Mindset
Self-Efficacy
Sense of Belonging
Relevance of School
Self-Awareness
Relationship Skills
Executive Functions
Attachment
Stress-Management
Self-Regulation

# Three Concepts that Go Beyond an Exclusive Focus on Average

2. ***Context***: The attributes shown by an individual at any point in time are shaped by the ***specific*** context of development.

# Context

- Monozygotic rodents exposed to different *in utero* and post-natal experiences

Slavich & Cole (2013, p.332)

# Three Concepts that Go Beyond an Exclusive Focus on Average

3. **Pathways**: We all walk the road less traveled: Every individual will have his/her own ***specific*** history of development across time and place.

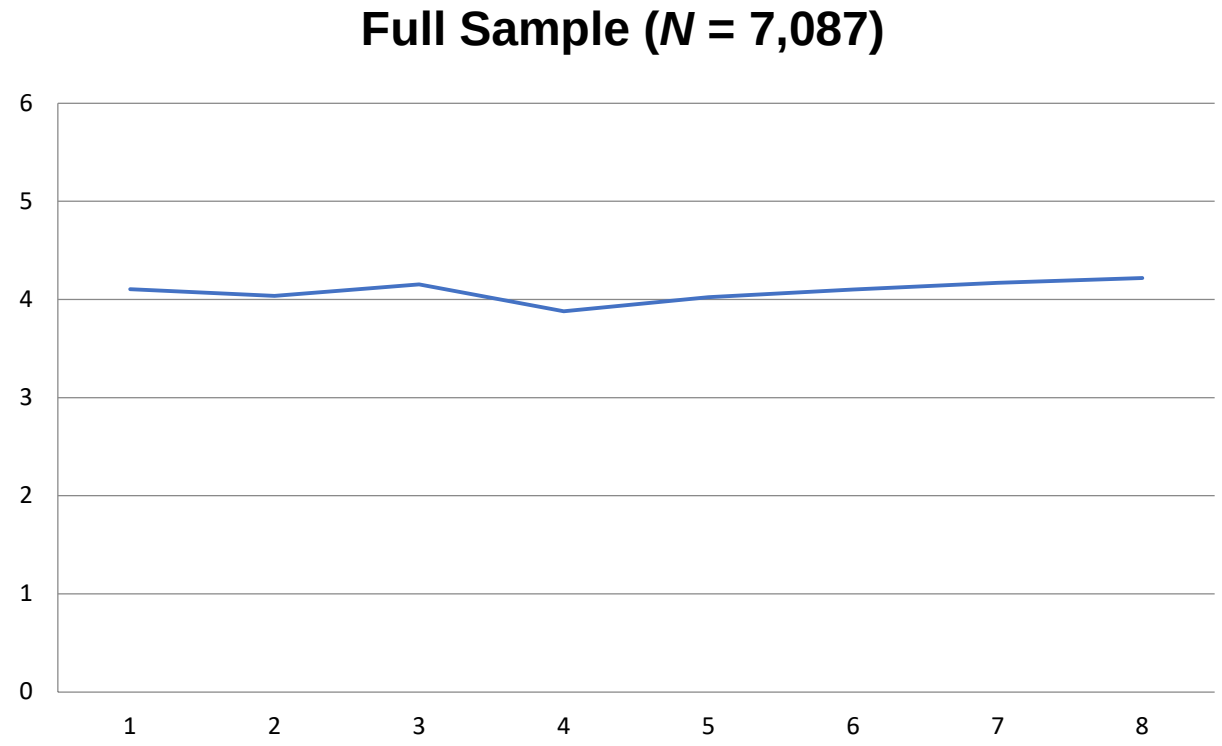
# Pathways:

## 4-H Study of Positive Youth Development

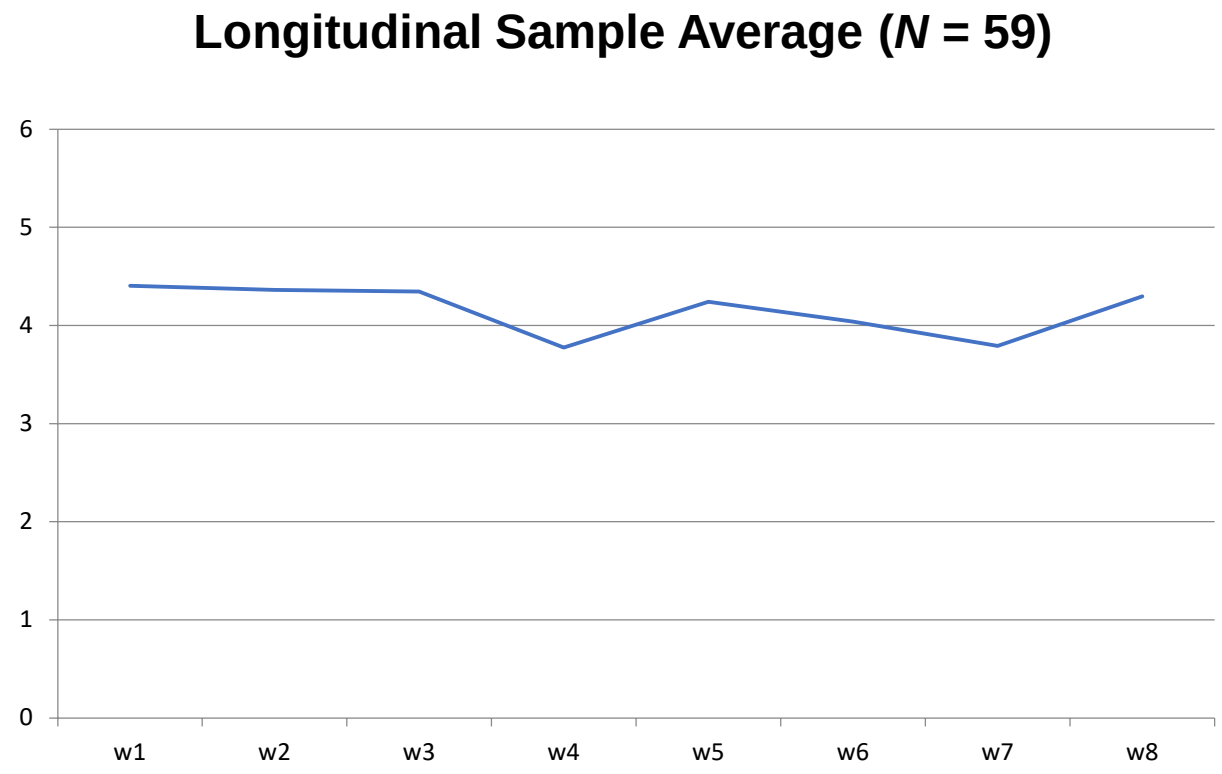
- 8 waves (Grades 5 to 12, respectively), 7087 participants
- Gender: 60.6% Female; 39.4% Male
- Race: 70.7% White; 10.2% Hispanic; 7.9% Black; 3.7% Multiethnic; 3.3% Other; 2.3% Native American; 2.1% Asian
- Mother's education: 33.6% 4-year degree or higher; 37.2% 2-year or technical degree; 20.5% High School; 8.6% less than High School
- Mean per capita income $15,279.26

*Note that the longitudinal sample presented in these analyses includes all cases with at least 6 waves of data.*

# Goal Optimization Skills:
(e.g., strategic thinking, executive functioning, resource recruitment)

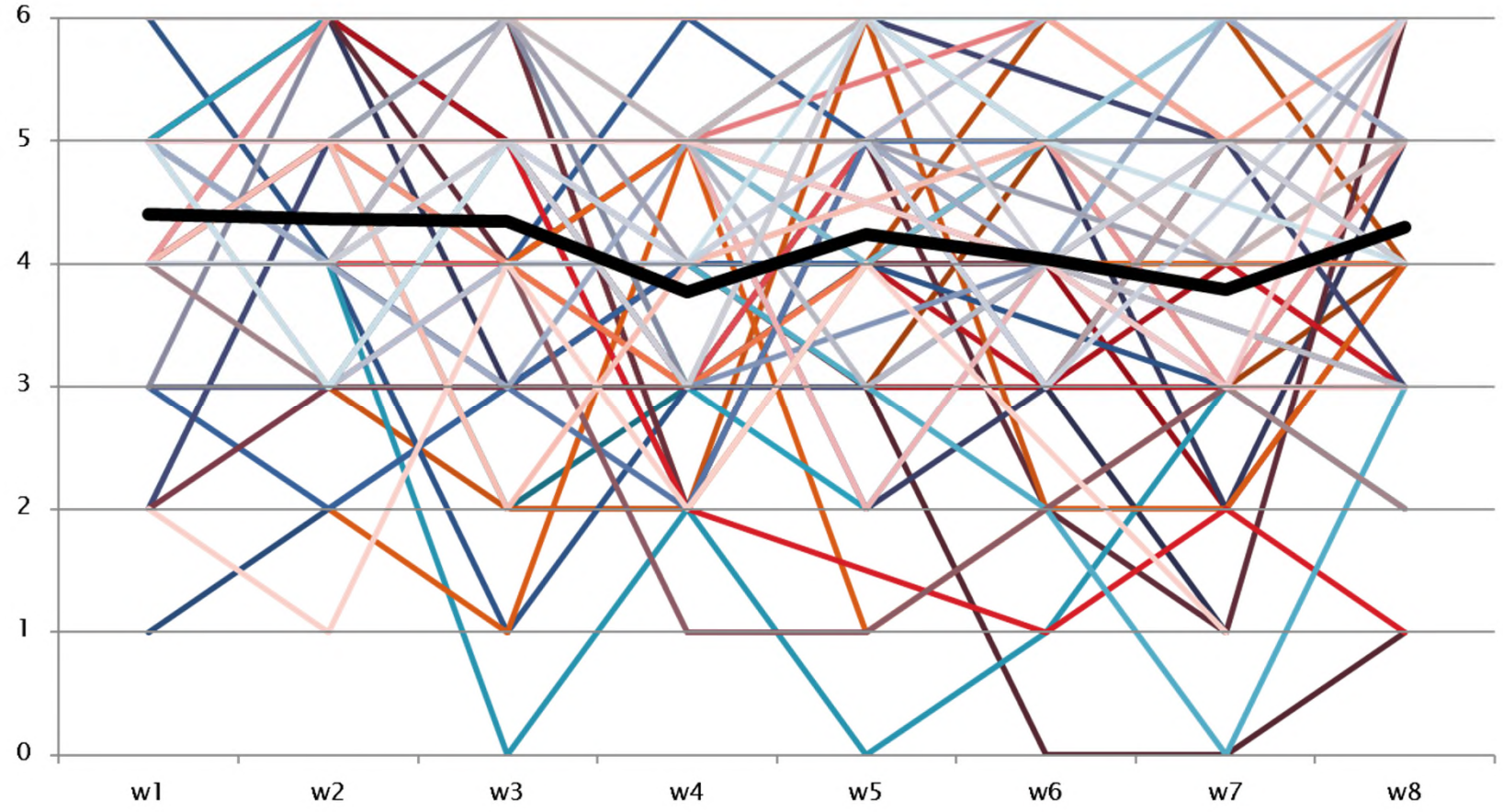


# Goal Optimization Skills:



# Goal Optimization Skills:

## Person-Specific Pathways for Longitudinal Sample (*N* = 59)



# The Bornstein Specificity Principle

**Development Involves Relations Between a Specific Individual Occurring at a Specific Time and in a Specific Place:**

- The *Specificity Principle* emphasizes that specific contextual conditions, of specific people, occurring at specific times, shapes specific facets of development (e.g., physiological, psychological, sociocultural) through specific processes of individual ⇔ context coaction.
- The Specificity Principle applied to parenting: Specific experiences that specific parents provide to their specific children at specific times in the lives of children have specific effects on specific aspects of the children's development.
- The Specificity Principle emphasizes the distinctiveness of each individual's development.
- ***Implications for program evaluation: Evaluation must begin with a focus on changes within specific people, and not on relations among variables across people!***

# A Prototypic Use of the Bornstein Specificity Principle to Frame A Person-Specific Approach to Program Evaluation

- Does a specific treatment,
- Of a specific duration and intensity of dosage,
- Within a specific program design,
- Promote a specific degree of change,
- In a specific variable (e.g., the character virtue of humility)
- For a specific program participant,
- At a specific point in development,
- From a specific family, community, society, and culture,
- And at what specific time in history?

**THERE ARE METHODS THAT CAN ADDRESS THIS COMPLEX QUESTION**

# Peter Molenaar and John Nesselroade: Human Development is "Non-Ergodic"

- Every human has characteristics that are possessed by all humans (e.g., respiratory systems or circulatory systems). These attributes are termed "nomothetic" attributes.
- Every human has characteristics that only some humans have (men have one reproductive system and women have another reproductive system). These attributes are termed "differential" attributes.
- Every human has characteristics that no other human has (e.g., each person has a specific genotype, composed of nuclear + mitochondrial DNA, and a unique epigenetic history). These attributes are termed "idiographic" attributes.
- Each person has a specific constellation of nomothetic, differential, and idiographic attributes.
- Therefore, the specificity of each person is assured.

# Peter Molenaar and John Nesselroade: Human Development is "Non-Ergodic"

- Molenaar, Nesselroade, and their colleagues (e.g., Nilam Ram, Sy-Miin Chow, Ellen Hamaker, Alexander von Eye, and Lars Bergman) have developed statistical tools to analyze the development of a specific person's (idiographic) development across a large number of times of testing (e.g., 50 to 100).
- For example, a procedure termed "dynamic factor analysis" (DFA) enables an individual's specific ("idiographic") pathways of change to be measured across a series of such intensively sampled time points.
- Unlike traditional "time-series" data-analysis methods, DFA enables earlier times in development to influence subsequent times.
- In turn, through another statistical procedure, termed "Idiographic Filtering," it is possible to determine if two or more individuals have similar (differential or nomothetic) pathways, at least at the latent-variable level.
- Therefore, by starting data analyses at the person-specific level, these methods can: 1. identify specific, meaningful individual (idiographic) features of development and, as well, 2. meaningful group (differential) or general (nomothetic) features of development.

# CONCLUSIONS

- Methodology in developmental science continues to advance in its usefulness for producing valid results in the enactment and evaluation of programs aimed at promoting character virtue development and human flourishing.
- Methodological advances are also increasing in complexity.
- Applied developmental scientists stand ready to collaborate with colleagues and communities to do their share of “heavy lifting” to enable programs to increase their capacities to promote flourishing.
- These collaborations hold the promise of better serving the specific goals for health, positive development, and lives of meaning and mattering for specific individuals in specific communities in specific countries around the globe.
- The series of webinars (of which this is the last) led by TWCF for their GICD Initiative has been a great example of this collaboration.
- We hope to continue to collaborate you.

# QUESTIONS, COMMENTS, AND THANK YOU FOR YOUR ATTENTION!