Final Report
Submitted to the Board of Developmental Research Institute
For grant-funded project:

Investigating the validity of the Ogive method, including the use of Rasch Analysis, for Sentence Completion Test assessment for the STAGES Model

Tom Murray
June, 2018

*** Internal Report -- Not for Distribution without permission ***

Executive Summary

This project had these goals:

- To further investigate problems with the Ogive cutoff method traditionally used to aggregate response scores for sentence completion tests in the Loevinger tradition.
- To suggest alternatives to the Ogive method.
- To investigate the use of Rasch analysis in assessing the psychometric validity of the STAGES model.

Outcomes summary:

- **Overall test strength**. The STAGES inventory and scale is psychometrically very sound across its entire range (with more certainty about this for the middle levels). Its strength had been verified in the past by repeatedly high Cronbach's-alpha measurement of internal consistency ("reliability"), but the IRT and Rasch analysis adds additional depth to its validity.
- **Item analysis**. All items on the general protocol are strong, except one stem: "I am..." Terri confirms that anecdotally this item has unique responses. This item would be the first to consider removing in a shorter test. However, it might point to measuring an important sub-trait, suggesting further research. All other items are strongly correlated with each other. Items are assumed to be of equal "difficulty" and this is

mostly true, but analysis shows which ones are in fact a bit more "difficult" on average.

- **Shorter tests**. The test overall is strong with even 10 or less items. This finding has supported Terri in creating short-length assessment products in recent moths. Also, the new Focal Score rating (see below) makes it easier to produce a total score for protocols of any length (one does not have to worry about inventing new cuttoffs).
- **Factor analysis**. As concluded by prior research on the WSUCT, the SCT loads on one factor, i.e. seems to be measuring one latent variable. I.E. among the 36 stems none seem to cluster together as if they were measuring a sub-skill of the overall capacity. (Note that even in analyzing the Specialty Protocols, the 6 specialty stems do not seem to cluster as a group compared with the other 30 'general' stems.)
- **Scale resolution**. Though the STAGES scale has 12 levels, scoring seems to be able to strongly differentiate/resolve only 6 levels across the scale. I.E. it has some difficulty in differentiating the difference between active vs. passive. Terri thinks this makes sense in terms of the gradualness of stage transitions.
- **Rejecting cutoffs**. Any cutoff method is problematic, for reasons explained below; and the Ogive method has additional problems explained in detail. Therefore we stepped away from using any type of cutoff method.
- **New aggregation methods**. Mathematically speaking, the most well-understood and reliable method of aggregating over item scores to produce a total score is the sum, or, equivalently, the average. However, this method contradicts with the intuition that, for projective instruments, many scores are "throw-aways" that should not lower the total, and that higher scores should receive more weight. After comparing various schemes, including weighting schemes, we decided on a method of taking the *average of the top x items*. We then decided on the top 6 items, as this resulted in the measurement closest to the prior Ogive scores (based on root-mean-square error). Finally, we made a slight linear adjustment to bring the new scoring system even closer to the prior Ogive scores, to support having the minimal confusion and effort in transitioning to a new method.
- **Final formula**. From the above considerations, the new method for scoring is (tentatively) called the ***Focal Score***, which is: **(1.2*Ave-top6)-1**. This new measure has been added to the STAGES Training Platform, and will seen be included in the Scoring Platform. For several months Terri has been observing that the Focal Score matches very well with her intuitions about stages, close to the old Ogive score, and even better than the old Ogive score in the Concrete tier.
- **Maturity vs. Shadow**. The old method is thought to conflate lower maturity with shadow, and the new method separates these factors. This

allows for better results of children, and gives both a center of gravity and a shadow indicator for adults.

- **Additional measurements**. The new scoring system interface shows the scorer several measures in addition to the Focal Score: The *Average* (across all stems); *Bottom Score* (the average of the bottom 6, related to a "shadow score" for adults); *Spread* — the spread score is the Focal Score minus the Bottom Score, and is an indication of the range of values in the survey.
- **Equal spacing of levels**. Rasch analysis shows that the levels defined by the STAGES model have relatively equal spacing, which is a tentative confirmation of a hypothesis claimed by O'Fallon.
- **Data set analysis**. This report also shows various other perspectives on the STAGES data. This includes a "Skew and Kurtosis" analysis showing that, for each person (each survey), the distribution of scores is normally distributed (a bell-curve) on average. This is a bit surprising because the conventional assumption might be that, for the populations in the database, there is a longer "tail" of off-the-cuff (or shadow) answers into the lower levels.

Background

Project beginnings

There were two initial motivations for this project. First, Terri's statisticians (Nayak and Moni) suspected some fundamental errors in the assumptions behind the Ogive cutoff method for aggregating over the (36) response scores producing the total survey score, as traditionally used in Loevinger-lineage sentence completion tests (SCT). One goal was to identify alternatives to the Ogive method.

Second, the main alternative framework to STAGES and other SCTs are the Lectica assessments (based on Hierarchical Complexity Theory, HCT, see Commons et al.; Dawson et al.; Fischer et al.), and research in that lineage uses Rasch models for evaluating psychometric validity. Rasch analysis is a subset of Item Response Theory (IRT) methods, and IRT methods represent a more robust and modern set of methods than the one's Loevinger chose over 30 years ago. A goal was to upgrade methods to more modern techniques, and also to see if Rasch analysis can shed light on how to replace the Ogive aggregation method.

Work for this project began informally in the Fall of 2017, following conversations between Tom and Terri. The official grant work proposal was submitted in September 2018, and a final report was given on a zoom call in December 2018. Additional follow-up work was done in January and February, and drafting this written report comes some months after completion of the project.

Initial consultations with IRT experts indicated that the SCT projective test has some unique properties that present interesting challenges. The Ogive cutoff method (see Appendix) essentially "throws out" data beyond the cutoff. This aligns with the general intuition that for a projective test, where subjects are not trying to "do well" or "get it right," that the highest scores should be given priority, indicating the actual competence of the subject. However, these considerations are unusual in the world of psychological testing, and none of our statisticians knew of other accepted methods that would meet these requirements (or needs). With no easy out-of-the-box solutions available a number of potential alternatives needed to be explored.

We expect that this project will not only situate the STAGES model on more sound psychometric ground as its use expands, but it will also produce results that will be of general interest to the entire SCT (Loevinger tradition) community, and probably for the greater community of developmental researchers and practitioners. We have plans to publish a summary of these results in an upcoming journal issue (Integral Review). (Note that the scope of this funded DRI project does not include work for a published article.)

Project consultants

Murray enlisted consultation from several experts in the course of this work. Bob Dolan is a psychometrician in the field of educational technology, has worked for major companies including Pearson Publishers, and is a personal friend of Tom's. Matt Johnson is on the mathematics faculty and Columbia University, working in "Psychometrics and Statistics Research," and is an expert in item response theory. He is a well respected international leader in the field of IRT, who was recruited for us by Dolan. Trevor Bond is one of the leading scientists in the Rasch Analysis community, author of *Applying the Rasch Model* , and is a leading expert in the application of Rasch theory to developmental psychology (he was also one of the primary consultants for Theo Dawson, the main psychometrician in the Hierarchical Complexity community).

For this project Murray procured the primary software package used in the Rasch community, WINSTEPS, and consulted with Mike Linacre, its creator and another leader in the Rasch community, several times. It was originally thought that all consultants would receive some honorarium, but in the end only Bond accepted an offer to be paid (moneys paid directly by Murray). In the final months of the project, Terri's consultant Moni Blazej Neradilek (Mountain-Whisper-Light Statistics) was also brought in for advice.

The STAGES model

Additional information about the STAGES model can be found in papers by O'Fallon at https://www.stagesinternational.com/papers/. Also, Murray's "Sentence

completion assessments for ego development, meaning-making, and wisdom maturity, including STAGES," appearing in Integral Leadership Review, (August 2017) is available at http://www.perspegrity.com/papers. Finally, our major research work in validating the STAGES model is described in "The Validation of a Novel Method for Assessing Developmental Stages of Meaning-Making Across the Life Span" by O'Fallon, Polissar, Neradilek, and Murray, is in submission to a journal.

The Ogive and TWS methods

Within the Loevinger tradition (including Cook-Greuter's MAP assessment and the STAGES inventory before this study) there are two established methods for aggregating the 36 sentence completion scores to obtain an overall score (sometimes called "center of gravity") for a survey. The most commonly used by far is a cutoff method called the Ogive method. In the Ogive method one checks each level in turn to see whether there are sufficient scores at or above that level. The cutoff number varies over the levels, as detailed in the Appendix 1. Cook-Greuter says the Ogive (or center of gravity score) "is assumed to reflect a person's most ego-syntonic and habitual frame of reference or preferred mode of responding to life."

The other aggregation method, which is usually calculated along with the Ogive, but is less frequently used, is the total weighted score (TWS) method. This method essentially sums the scores over the 36 items. The TWS method returns an integer, typically in the 300-600 range. The formula used for STAGES is: (f_1.0 * 2) + (f_1.5 * 3) + (f_2.0 * 4) +...+ (f_6.5 * 13), where, e.g. f_1.5 is the count (frequency) of completions scored at 1.5.[1] (There are slight variations on the constants used in this formula for different frameworks.)[2]

The Ogive method is preferred because it is believed that the categorical result is more meaningful and has more construct validity that the TWS method. The theoretical model, the scoring manual, and the scores assigned to individual completions are all based on a sequence of discrete categories. Researchers in the psychological and social sciences almost universally acknowledged that psychological phenomena is complex and that using discrete categories involves substantial simplification—and is done for practical reasons. There is in fact

---

[1] Note that the Ogive cutoff method, which returns a category (ordinal) value, must be implemented with a procedure (i.e. checking the cutoff at each level in sequence, and if it is not satisfied moving to the next level...), whereas the TWS method is implemented with an algebraic equation.

[2] The "weighted" in TWS is actually misleading, as the score is very close to a linear sum of the 36 scores. The term weighted is probably an artifact of the fact that originally the data was represented using frequency vectors, i.e., for a system ranging over 10 levels (as in MAP) the data had 10 values, each showing the count of scores at that value. The method was used prior to personal computers, and saving 10 numbers was easier than saving 36 numbers (the values for each completion). Unfortunately, because of this, archived data from most surveys in this tradition is missing important information. We know that, for example, a survey had 14 items at the "Expert" level, but we don't know *which* of the 36 items scored at the Expert level. This makes certain types of analysis, including IRT analysis, impossible with such archived data.

considerable debate within the developmental psychology community about whether development happens continuously or in spurts that create quasi-discrete levels or stages (see Dawson et al., 2005 for some evidence of discrete spurts). However, no research has tested whether the construct of ego development actually grows continuously or in spurts, so the use of categories rather than a continuous scale is for convenience purposes.

Loevinger's scoring manual does allow for assigning "borderline" scores; and the manuals developed by Cook-Greuter, Torbert, and O'Fallon, allow for the refinement of the stage-based scoring to specify whether an item is early, central, or late within a stage.

Though the discrete Ogive measurement is used predominantly, the continuous TWS measurement is sometimes used for comparative, pre-post, or longitudinal studies because it is more sensitive to small changes. (It can take years for individuals to change one developmental level, if they do at all.)

The Ogive cutoff values are, theoretically or conceptually, inspired by Bayesian probabilistic principles: i.e. how much info is needed to conclude a person is at level X with 95% confidence? (95% is usually chosen as the confidence factor to use.)

As noted in Appendix 1, the cutoff procedure starts with the top level, and proceeds to check each lower level until 3.0, at which point it *switches* to testing the very lowest level and works is way *up* to 2.5 (the same method is used for Loevinger and Cook-Greuter, with slightly different stage sequences vs. STAGES). This follows the principle of checking for the least likely possibilities first. The 2.5 level ("Conformist" in earlier models) was determined to be the "average" or most frequent level, and thus the final step of the procedure selects 2.5 as the default "if there is not sufficient evidence for any of the other levels." It is the "default" level, i.e. the one we would guess given no other information (i.e. no test).

Problems with Ogive and TWS aggregation methods

Here we arrive at the purpose of our present study. The Ogive method has several drawbacks, and even seems to include some errors within the assumptions used to determine the cutoff numbers. Surprisingly, given the large number of studies using the SCT, the Ogive method for accumulating items is rarely questioned or analyzed. Researchers use the cutoff's given by Loevinger without question (the vast majority of the 400-plus published studies on the SCT are based on Loevinger's version of the SCT). One reason for this may actually be the large number of studies in the field—one wants to build upon and be able to compare results with prior studies, and questioning foundational assumptions of a well-established framework might appear counterproductive to many research goals.

We will mention problems with cutoff methods in general, and also problems specific to the Ogive cutoff method:

1. Lack of empirical support for discontinuous levels
2. No population studies for Bayesian estimates.
3. Difficulties estimating extreme values
4. Lower cutoffs confuse development with "shadow" evidence
5. Sensitivity to error
6. TWS misalignment
7. Additional issues with the TWS multipliers
8. Errors in the Ogive formula and assumptions

Lack of empirical support for discontinuous levels. The cutoff method forces survey scores into discrete categories, but there is no evidence that ego development grows in discontinuous spurts. One of the leading researchers using the SCT notes: "The Ogive rules are ingenious procrustean devices to make gradual transitions look like phase changes, translating continuity into discontinuous types. Naturally, therefore, at the borderlines, scoring gets much more unreliable... Furthermore, since these rules put special emphasis on the few most extreme scores, [the total scores] become even less reliable the farther we get from the center of the distribution of the population" Holt (1980, p. 918).

No population studies for Bayesian estimates. A serious problem with the cutoff method is that, based on Bayesian probability concepts, it makes assumptions about the overall population. There are three problems here.

(1) In fact there are no large scale (population-scale) studies using the SCT (in part because it is time-consuming and thus expensive to score). Loevinger's estimate of Conscientious as being the average (and default assumption) was an educated guess based on a number of studies (each with subjects in the hundreds at most). But many of these studies were with populations such as pregnant mothers, prison populations, and others with below-professional educational or socioeconomic status. This may in fact be a reasonable guess for the average US population, but the modern uses for the SCT (e.g. by Cook-Grueter, Torbert, and O'Fallon and their affiliates) are predominantly within professional or self-improvement contexts, i.e. where "leadership skills" and "self reflection skills" are considered important. In fact, articles by Cook-Greuter and O'Fallon have tried to estimate the percentages of each developmental level within specific populations, e.g. general population, college-educated, professionals (mangers), and consultants (see Cook-Greuter 2002; Torbert & Livne-Tarandach, 2009).

(2) Within any specific population, the base estimates ("prior probabilities") will differ, and thus the Bayesian assumptions differ, and thus the cutoffs should, ideally, be modified. But this is never done. In practice the lower and middle cutoff numbers, including the procedural "switch-over" at Expert and the "default" value at Conscientious (2.5) are never modified. This is because, first, one wants results to use standard methods and to be comparable with past research in the field, and

second, gathering enough data in any sub-population to determine the proper new cutoffs would usually be prohibitively difficult anyways.

(3) Several sources indicate appears that the population as a whole becomes more sophisticated over time, suggesting that the base rates should evolve over the decades. This factor is also ignored in keeping with standard Ogive cutoffs.

Difficulties estimating extreme values. The Bayes-based method for determining the Ogive cutoffs (details in Appendix 1) is particularly problematic near the upper and lower extremes, i.e. for the rarest scores. Loevinger says "The rule is not helpful at extreme ratings. However, the number of cases per category at the extremes is so small that theory or intuition is a more reliable guide" (pg. 121). Cook-Greuter says "employing Bayes theorem for the problem of determining cut-off numbers at the high-end was a creative move by Loevinger at the original time...On the other hand, I found Bayesian reasoning alone inappropriate at the upper end...Therefore, I felt that cut-off numbers needed to be related to the actual number of post-autonomous responses observed...rather than merely based on probabilistic calculations" (1999 dissertation, p. 228). Cook-Greuter chose stricter cutoff criteria for the top levels vs. Loevinger. Cook-Greuter notes, that the question of "how many pieces of evidence one requires to draw a conclusion, is a necessary step. It is also arbitrary—arbitrary not in the sense of random or capricious, but in the sense of arbitrated, reasoned, deliberated" (p. 224).

The populations of interest in many contemporary assessment projects are actually in or near what Loevinger considered the "extreme" of the upper end. Thus these problems are front and central to modern assessments, rather than being relinquished to rare occurrences. Apparently recently Cook-Greuter and associates have adjusted their cutoff values and no longer rely on Bayesian reasoning (personal communication). They do continue to use a similar cutoff scheme, which is thus expected to have most of the issues mentioned here. Currently there is no published material explaining the reasoning or values for a newer method of determining cutoffs. We expect that any new method is informed by reasonable statistical principles, yet still admits to a large degree of arbitrariness and/or setting key parameters based on the intuitions of a highly experienced expert (Cook-Greuter).

Lower cutoffs confuse development with "shadow" evidence. The scoring rules, including the examples given in the scoring manuals used for WUSCT and MAP, seem to confuse immature development of ego with "shadow-crashes," i.e. momentary regressions to lower ego levels that may be triggered based on the subject of the stem. Because the levels 1.0 and 1.5 are "rare" in normal adult populations, they are checked before 2.0 and 2.5 in the ogive rules (see Appendix 1). A person needs only 7 of the 36 stems rated at 1.0, 1.5, or 2.0 to "be at" those levels, respectively, in their total score (Ogive). Getting such a low rating can come from a very simplistic one-word answer, but it can also come from what seems a narcissistic or impulsive answer, including the use of vulgarity or mentioning of

violence. I have made the case that the scoring method confuses actual level of development with moments of such "shadow crashes."

For example, a person may have 19 stems at 3.5 (Achiever or even higher), yet will get an Ogive score of 2.0 (Rule-based) if they have only 7 stems at 2.0 or lower. Clearly such a person's "center of gravity" should be more like 3.5 than 2.0. Because there are very few people scoring below 2.5 in the populations (or business clients) that currently use the SCT, this problem has not been particularly salient. One area where it is limiting is in studying children. O'Fallon wants to use her model (and in fact already has some data) for pre-adult populations. In such cases one wants a more reliable measure of ego development that is not conflating emotional outbursts or "triggered" responses with simply lower levels of ego development (perspective taking).

These problems of conflating maturity with shadow in the lower levels is an additional reason for moving away from the cutoff system. It suggests having separate measurements for developmental level (e.g. an sum or average score like the TWS) vs. a shadow-related measurement (e.g. counting the number of low-level impulsive scores).

Sensitivity to error. As mentioned by Holt, surveys with score counts near the cutoff levels are particularly sensitive to error introduced by random or uncontrolled factors, including scorer variability. E.G. if the cutoff for some level is 10 items, and a survey has 9 or 10 or 11 items at that level, then a small difference in item scores, say due to scorer error or to a person not expressing themself "at their best" in the moment of completing the stem, would have resulted in bumping the total score to a higher or lower developmental level. Each developmental level has a significantly different description than its neighbors, so this small difference can create a large change in outcome. (O'Fallon often scores to an extra level of detail, specifying early, mid, or late stage, but most trained scorers do not do this.)

TWS/Ogive misalignment. The cutoff method produces unsystematic alignment with the TWS (or any linear item aggregation), making them difficult to compare. See Figure 1. For example, for at TWS=260 (red horizontal line) there are three possible Ogive values; and for an Ogive value of 3.5, the TWS scores range from approximately 210 to 265. Since the TWS is basically a straightforward linear sum (see above footnote) of the scores, it corresponds to a more traditional and scientifically understood measurement. Scoring manuals include rough advice on how to compare TWS and Ogive scores. Loevinger (1998, p. 26) includes a table with TWS ranges, e.g.: E4 conformist > TWS 146-162; E5 Self-aware > TWS 163-180; E6 Conscientious > TWS 181-200. It is problematic that the Ogive has such a complicated mathematical relationship with the more basic and intuitive summation statistic.

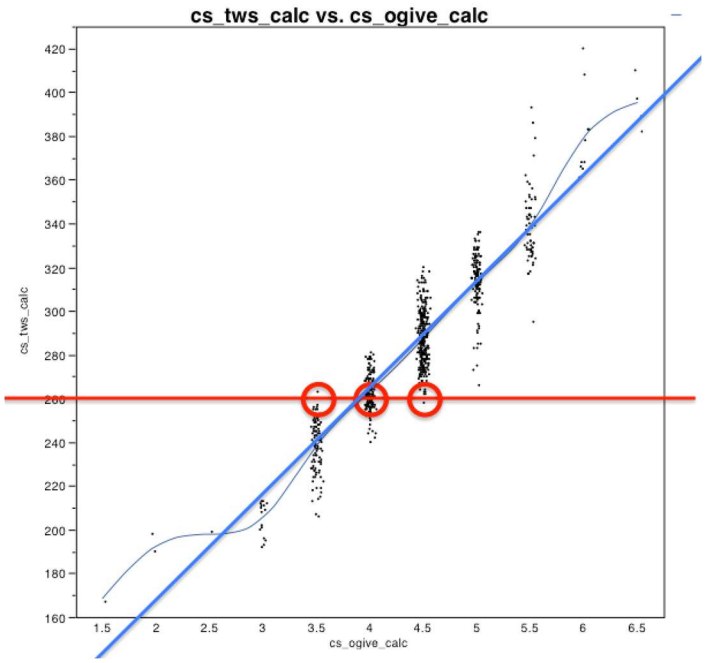


Figure 1: Ogive vs TWS

Additional issues with the TWS multipliers. The progression of weights used for the TWS calculation is somewhat arbitrary. In the MAP system the lowest frequency is multiplied by 2, and the next by 3, etc., with the multiplier increasing by one per level (see Appendix 1). This assumes an equal separation between each level. However, because Loevinger (and later Cook-Greuter and Torbert) based the model on intuitions from raw data rather than on an underling theoretical model (as STAGES does), the number of levels, or the way the "pie" of the sequence was cut up, was modified several times over the course of years of research. There was nothing in the theory or data analysis that could answer the question of equal spacing between the levels, and thus the TWS multipliers remained arbitrary, and kept shifting. With the insertion or collapsing of a level, the TWS multipliers are re-set. For example, they might still increment as in 2,3,4..., but these multipliers could correspond with different levels in any revised system. This made revised TWS values incompatible with prior results. Our new system replaces the ogive cutoff with a system that uses sums/averages, but does so in a way that does not have arbitrary multipliers, and will not have to be modified as the model/theory evovles.

Errors in the Ogive formula and assumptions. Finally, several of our statisticians (Dolan, Johnson, Polister, and Blazej-Neradiek) have pointed to fundamental flaws in the reasoning behind the equations cited by Loevinger and Cook-Greuter. It would appear that of the many hundreds of studies building on Loevinger's work, none of them checked these basic assumptions. We believe that the flaws in these assumptions indicate only small mathematical errors on average, and do not imply the need for significant revisions in past studies, but, desiring a strong foundation for the SCT, these errors are another reason to move beyond the Ogive cutoff method. The most effected conclusions of past studies are probably those that make

inferences vs. the general population. Those types of inferences are a very small percentage of the results of published studies, which usually are looking for correlations between ego development and other phenomena (conclusions which we believe remain largely valid in light of the errors we will mention), rather than making claims about whole populations.

Consultants Nayak and Moni believe that the Ogive formulas: (1) tries to assume that the survey items are independent, which they are not; (2) makes an error in the probability algebra; and (3) that the parameters (Prev, TrPos, FaPos in equation below) are too arbitrary and sensitive to assumptions:

$$\frac{\boldsymbol{Prev} * \boldsymbol{T_p}^{\boldsymbol{x}}}{\boldsymbol{Prev} * \boldsymbol{T_p}^{\boldsymbol{x}} + (\mathbf{1} - \boldsymbol{Prev}) * \boldsymbol{F_p}^{\boldsymbol{x}}}$$

Though the Ogive method is inspired by Bayes theorem, it also seems to violate certain principles in Bayes theory. In the Ogive procedure (Appendix 1), if a cutoff values is satisfied (e.g. 4 items at or above 5.5) then all other information is disregarded (i.e. the other 32 stems). But central to Bayesian theory is the idea that each new piece of information accumulates to inform the estimate of the resultant probability.

We do not have space or time here to flesh out or prove all of the issues and errors that the experts suspected at their first glance. Johnson said that the method "Makes sense only if the number of people above and below a level are more or less equal, which is not usually the case."

That there seemed to be grave problems with the ogive method was agreed upon by *all* of our consultants who familiarized themselves with the equations used (Dolan, Johnson, Neradilek , and Polisar). In addition, one of our Rasch consultants, who took a cursory look at the ogive method, called it outright "indefensible."

Item Response Theory

Rasch is usually seen as a subset of IRT that, for explicit theoretical reasons, is limited to "one parameter" modeling (IRT includes a family of 1, 2, and 3-parameter models). IRT (also known as latent trait theory, strong true score theory, or modern cognitive test theory) is one of the main schools/toolboxes in modern psychometrics (it is the method used in validating high-stakes tests such as the SAT and GRE). It specifies statistical methods for use in assessments that include many items to assess a single human trait (which is called the "latent variable" because it can not be measured directly).

IRT creates a single scale that is used to measure the latent variable (i.e. the human skill or "proficiency") *and* the difficulty of an item (and the score for the whole test). In this way item difficulties and person abilities can be meaningfully compared. For example, a value of "3.2" means the same thing, in terms of the latent variable, when

used to talk about the skill level of an individual, the difficulty level of an item, or the overall score on a test. In IRT a test does not have a single "reliability." Instead, a test might have stronger psychometric strength at some trait levels vs. others.

An item's difficulty is defined as the trait level required for participants to have a 50% probability of answering the item correctly. If an item has a difficulty of 0, then an individual with an average trait level (i.e. of 0) will have a 50/50 chance of correctly answering the item.

By using "logistic" scaling, the latent scale is always normalized such that (1) the mean value is 0 (i.e. negative values are below average); (2) the range is a true scalar (continuous or interval scale) in that it is infinite in positive and negative directions (even if the test items vary over a specific range, as in a 1-5 Likert scale or a 0-100% correct score for a math problem); (3) the standard deviation is 1 (i.e. the scale works conceptually like a bell curve or "Z-score" with standard deviations marked on the x axis).

One thing that distinguishes IRT from the "classical" test theories that came before it is that it does not assume that each item is equally difficult—i.e. it calculates an item difficulty for each item. IRT assumes that the probability of a correct response to an item is a mathematical function of person's skill and the parameters (primarily difficulty) of an item. IRT analysis starts with a sufficiently large set of test data and, using that assumption, works backwards to calculate (1) the difficulties of each items, and (2) the overall skill of each test-taker. IRT models posit that the difference between an item's difficulty and a person's ability reflects the probability of a person succeeding on a given task (at least for the 1-parameter version).

One strength of IRT (and Rasch) analysis is that the parameters of the models are generally not sample- or test-dependent, and can thus be generalized to different contexts. (Note that "polytymous" IRT methods were used here, i.e. those applying to non-binary measurements.)

IRT analysis can be used to answer questions including: do any of the items *or* any of the test-takers seem to be *outliers*; i.e. have an unusual relationship between skill level and item response? Does the test as a whole comply with the assumption that it measures a single latent variable (i.e. is it uni-dimentional)? For more information, see the Figure in Appendix 4 "IRT vs. Classical test theory."

Note that the SCT, a projective test, theoretically assumes that all items are equally "difficult" on average. The stems are intended to "triangulate" from many life perspectives, such that for a given domain (family, self, work, public life, etc.). Though individual may have some more and some less mature responses, the differences are assumed to be entirely personal, and not based on general item difficulty. We can use IRT to test that assumption.

Rasch Analysis

Rasch analysis (Rasch, 1980; Bond & Fox, 2001) is usually considered a subset of Item Response Theory, though die-hard Rasch proponents maintain that its underlying philosophy of measurement theory is different. IRT includes 1, 2, and 3 parameter models—while Rasch advocates advocate for no more than 1 parameter. "Parameters" are numbers that characterize each item differently. The first and primary parameter is the item difficulty. The second parameter is the item "sensitivity," which measures essentially how quickly the item responds to changes in skill level. A maximally sensitive item will have a sudden jump between level X and the next level, with low and high plateaus on either side of that jump, meaning that it discriminates above and below that value (its difficulty value) extremely well, but does not discriminate among the lower values, and does not discriminate among the higher values. One could think of this as asking "how much *information* does an item contribute to differentiating each level from others?" The third parameter sometimes introduced is the likelihood of "guessing" the correct answer for each item.

For reasons related to foundational theories of measurement, the Rasch community says that only the 1 parameter model produces a "true" continuous scale of the type used in measurements in the hard sciences. They constrain analysis to use the 1 parameter model, and identify items (and subjects) for which the data fall outside of this model as outliers. In general this means the test item should be removed or revised. In contrast, the more standard IRT method takes a purely statistical modeling approach that uses as many parameters as needed to find a model that best fits of all the data. The details of this controversy out outside our scope, but our consultants have indicated that for our data set, the difference does not matter, and the 1 parameter analysis, i.e. compatible with Rasch assumptions, is sufficient.

Commons & Pekker (2009, p. 13-14) summarize the Rasch model saying that it: "Transforms raw data into unidimensional, abstract, linear, equal-interval scales if the data fit the model. Equality of intervals is achieved through log transformations of raw data odds. The Rasch model is the only model that provides the necessary objectivity for the construction of a scale that is separable from the distribution of the attribute in the persons it measures...It works by using probabilistic equations to convert raw ratings of items into scales that have equal intervals if the data fit the model. Such a scale can then be used as a type of objective ruler against which to measure the data on items as well as on respondents (Andrich, 1988). Statistically speaking, this scale will be linear."

Again, we used Rash analysis to answer the questions:

- "Is it a **strong** Test?"
- Are score categories **equally spaced**?
- Are any **items** weak?
- What are the relative item **difficulties**?

- What is the **resolution** of the test? (how many levels does it seem to differentiate?)

Research Results

Data sources and descriptive statistics

Tom has access to the STAGES assessment database, which at the time of this work had 1941 completed and scored assessments (each with 36 stem completions). The focus was on the General protocol, which had 740 surveys—this was the data set used for most of this study. (Of the 1941, 946 are older MAP surveys scored using the Cook-Greuter method before STAGES was developed. The database also includes over 200 Specialty Protocol surveys.) The 740 scores were used for the IRT analysis and Descriptive Statistics, but for the Rasch analysis the set was reduce to only those scored by the primary scorer (O'Fallon), n= 448.

As an ancillary outcome of this research, as part of data cleanup and preparation, Tom identified some errors, redundancies, inconsistencies, and other 'garbage,' in the database that were later cleaned up with the help of Dee. Thus one outcome of this project is that the database is in better shape for future research studies.

Appendix 2 shows descriptive statistics and charts summarizing this data set. The notes on these charts include:

Figure A2.1: Histograms of scores per level: A: per response; B: per protocol

- Overall there are many more protocols in the database at 4.5, showing a skew toward that level in STAGES scored clients.
- Note the curious dip at 4.0 for response scores. Its a bit of a mystery why there are so fewer 4.0 item scores.

Figure A2.2: Comparing protocol scores for STAGES vs. SCTi (MAP) scores in the database

- This Figure just illustrates that the older data in the STAGES platform database, which was scored using the older MAP method, has a different spread of scores vs. the more current STAGES scores. It points to the potential to compare STAGES vs. MAP data in future research questions.

Figure A2.3: Histograms of responses for each stem separately

- This Figure illustrates that there may be interesting differences in how each stem elicits scores from different levels; suggesting the possibility for further analysis at the stem level.

- Of particular interest are patters that seem to alternate every other level, corresponding to active vs. passive biases.

Figure A2.4: Per-stem histograms, with normalized version of the left side

- This Figure shows per-stem histograms as in the prior Figure, but adds normalized versions that highlight how each stem differs from other stems at each STAGES level. Again, Suggesting the possibility for further analysis at the stem level.

Figure A2.5: Item correlations.

- This Figure shows a colored "heat map" of the correlation between every item with every other item. Items are reasonably well correlated, with correlations ranging from .32 to .53.

Other data set analysis. We looked at other metrics describing the data set. This included a "Skew and Kurtosis" analysis checking whether the frequencies of levels for each person (each survey) were normal (bell shaped). Figure 2 illustrates how skew and kurtosis represent deviations from a normal distribution. Skew can be described as "lack of symmetry around the mean" and kurtosis can be described as having a long (heavy) tail. Acceptable values for both are considered to be within -2 to +2.

For our surveys the mean kurtosis was 0.37 and the mean skew was 0.18, i.e. on average the scores within a survey are normally distributed around the mean. This is a bit surprising because the conventional assumption might be that, for the populations in the database, there is a longer "tail" of off-the-cuff (or shadow) answers into the lower levels. That the score distributions within surveys tend to be normally distributed gave us confidence that more standard aggregation methods would be applicable.

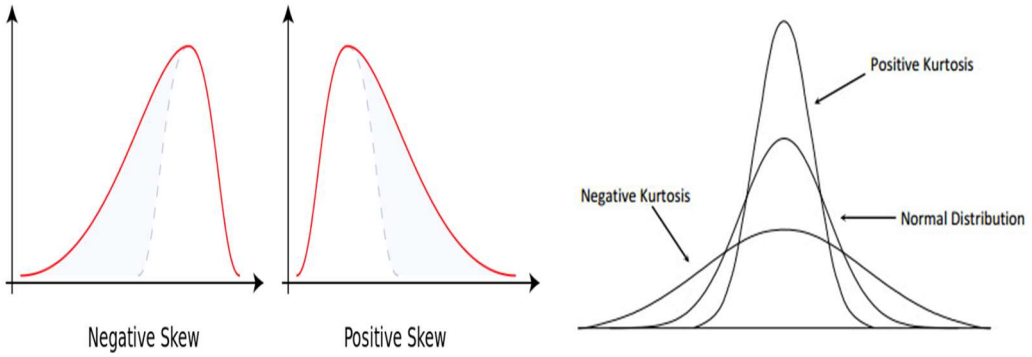


Figure 2: Illustration of Skew and Kurtosis deviations from a normal distribution

We also looked at how the standard deviation of scores within a survey varied based on the overall survey score (ogive). Results in Figure 3, shows that the standard

deviations for all of the levels (other than the top and bottom, 1.0 and 6.5) are remarkably similar (.8 to 1.2). (The error bars in the Figure show the variance in the standard deviations). Thus, overall the shape of the distribution of the 36 scores within a survey does not change much for individuals at different centers of gravity.

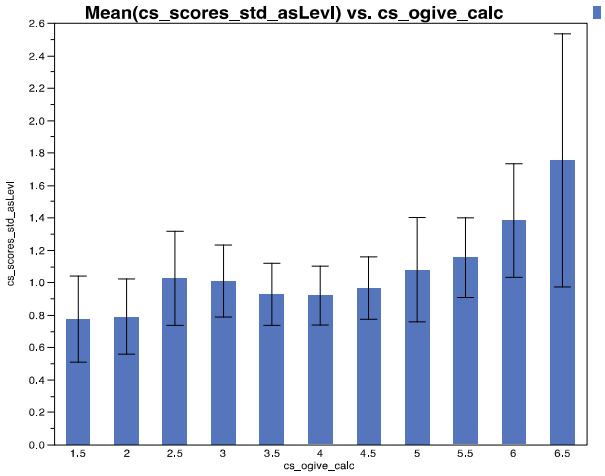


Figure 3: Standard deviation protocol scores vs. stage

IRT analysis results

One of the early conclusions was that Rasch and IRT analysis was not designed to provide a fix or straightforward alternative to the cutoff method (i.e. to provide a better method for determining the cutoff values). This is because it does not consider alternative methods for *aggregating* among items, it is concerned with the quality of the items *individually*, and it assumes that the overall score is a simple sum or average of the items (with variations in item difficulty taken into account). Rather, Rasch and IRT analysis provided insights about the quality of the test and specific items.

The following analysis was done using the R statistical package's most popular IRT module (the lmt R library), in consultation with our IRT consultants. Matt tried a number of alternative modeling methods on our data. He tried the 1, 2, and 3 parameter variations of IRT, and found that the 1-parameter version (which the Rasch model uses) was sufficient. Within the IRT analysis he tried the Generalized Partial Credit (GPC) model and the Graded Response Model (CPC), finding that the simpler GPC model was sufficient (both the AIC and BIC variations of the GPC model were tried, with little difference found). He also did some preliminary experiments using models outside of the IRT framework, including: "Mixture models" and "Grade-of-membership models"— and did not find evidence that these more complex methods added value to our analysis.

We can draw the following conclusions from our IRT analysis (all of which were verified in subsequent Rasch analysis):

Overall test strength. The STAGES inventory and scale is psychometrically *very sound across its entire range* (with more certainty about this for the middle levels). Its strength had been verified in the past by repeatedly high Cronbach's-alpha measurement of internal consistency ("reliability"), but the IRT and Rasch analysis adds additional depth to this finding.

Item analysis. All items on the general protocol are strong, except one stem: "I am..." All items have *good discrimination* overall —see the "test Information Curve" in Figure 3. This is an average of the test information over all 36 items. Figure 4 shows the item information for each stem separately. Though there is a dip for the middle values, all of the values are quite acceptable. Scoring is very *consistent* (i.e. good quality) across the prompts (similar to the Cronbach's-alpha results). It would be possible to do a deeper analysis on how each item performs across the range of all levels.

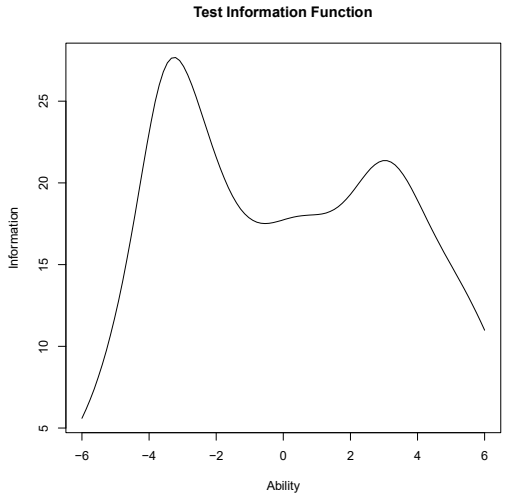


Figure 3. Test Information Curve.

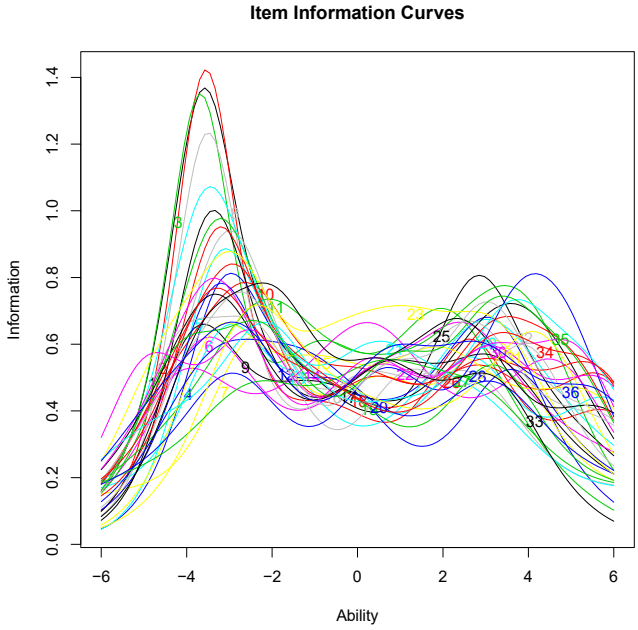


Figure 4. Item information curve, showing test information for each of the 36 stems.

Figure 5 shows the "Item Response Curves for two of the 12 levels"—illustrating the variability in how stems contribute to each level on average across the database. Figure 6 "Item Category Curves for two stems (ID 1, 9)" illustrates the same information from a different perspective, with different graphs per stem. In part these point to the possibility of more in-depth analysis of how each stem elicits responses from each stage.

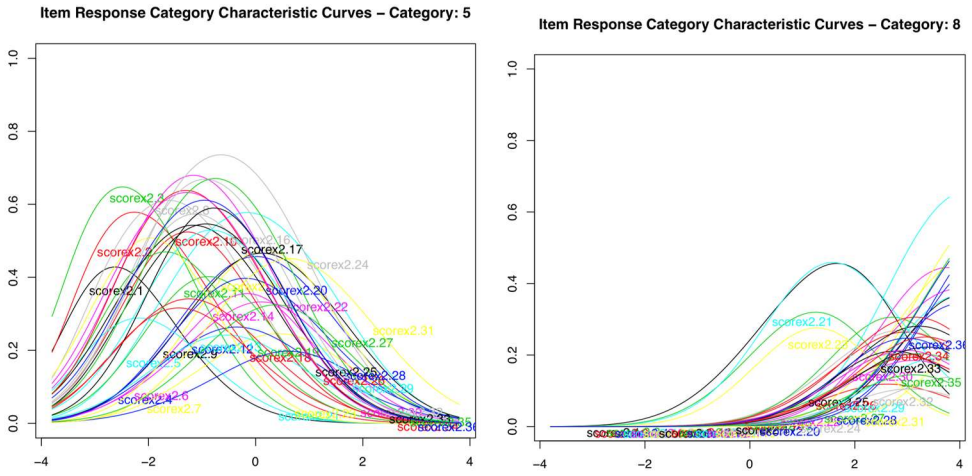


Figure 5: Item Response Curves for two of the 12 levels:
36 curves for each of "category 5" = 3.0 and "category 8" = 4.5

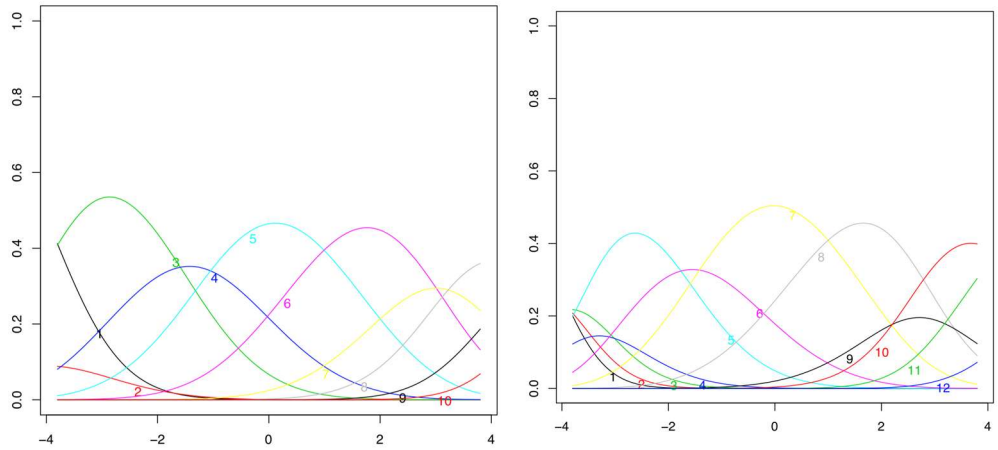
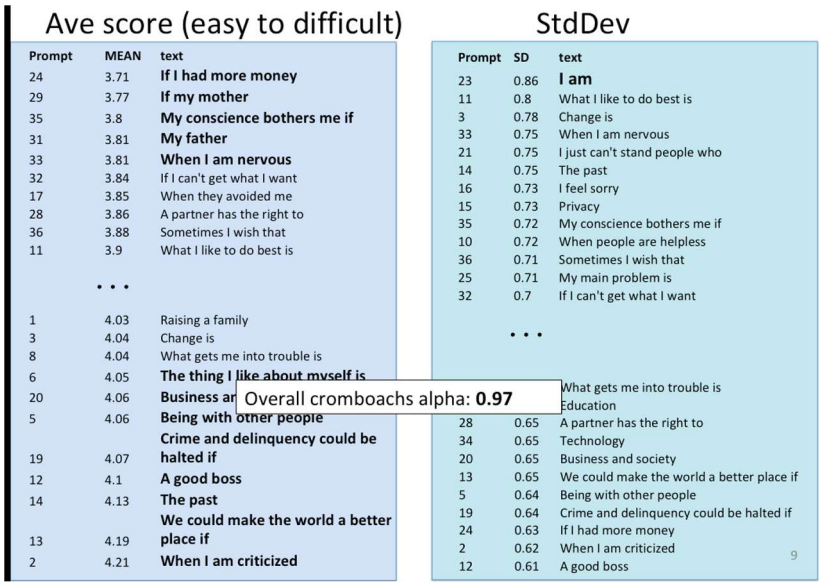


Figure 6. Item Category Curves for two of the 36 stems (IDs 1, 9):
12 curves, one for each STAGES level

The "I am" Stem: As mentioned it is the least well correlated with others and the overall "latent variable" (ego development) being measured. Terri confirms that anecdotally the item has unique and interesting responses. This item would be the first to consider removing in a shorter test. However, it might point to measuring an important sub-trait, suggesting further research.



Prompt	MEAN	text
24	3.71	**If I had more money**
29	3.77	**If my mother**
35	3.8	**My conscience bothers me if**
31	3.81	**My father**
33	3.81	**When I am nervous**
32	3.84	If I can't get what I want
17	3.85	When they avoided me
28	3.86	A partner has the right to
36	3.88	Sometimes I wish that
11	3.9	What I like to do best is
	• • •	
1	4.03	Raising a family
3	4.04	Change is
8	4.04	What gets me into trouble is
6	4.05	**The thing I like about myself is**
20	4.06	**Business ar**
5	4.06	**Being with other people**
19	4.07	**Crime and delinquency could be halted if**
12	4.1	**A good boss**
14	4.13	**The past**
13	4.19	**We could make the world a better place if**
2	4.21	**When I am criticized**

Prompt	SD	text
23	0.86	**I am**
11	0.8	What I like to do best is
3	0.78	Change is
33	0.75	When I am nervous
21	0.75	I just can't stand people who
14	0.75	The past
16	0.73	I feel sorry
15	0.73	Privacy
35	0.72	My conscience bothers me if
10	0.72	When people are helpless
36	0.71	Sometimes I wish that
25	0.71	My main problem is
32	0.7	If I can't get what I want
	• • •	
		What gets me into trouble is
		Education
28	0.65	A partner has the right to
34	0.65	Technology
20	0.65	Business and society
13	0.65	We could make the world a better place if
5	0.64	Being with other people
19	0.64	Crime and delinquency could be halted if
24	0.63	If I had more money
2	0.62	When I am criticized
12	0.61	A good boss

Figure 7. High and Low Averages and Standard Deviations for Stems

Item Difficulties and Standard deviations. Figure 7 shows "High and Low Averages and Standard Deviations for Stems" the General protocol. The range of average scores 3.71 to 4.21, a rather modest range, indicating a small variability in stem "difficulty." It is interesting to note that some of the stem difficulties correspond to what might be expect, e.g. "if my mother" and "when I get nervous" elicited lower scores, and "Crime and delinquency could be halted if" and "we could make the world a better place if" elicited relatively high scores on average. Figure 7

also shows that the stem "I am" has the highest standard deviation, consistent with the deeper IRT analysis showing it does not track with other stems particularly well.

Shorter test lengths. The test overall is strong with even 10 or less items. This has supported Terri in creating short-length assessment products in recent moths. Figure 8 shows a "Crombach-Mesbah curve" illustrating the Crombach's-alpha score for successively fewer stems (with the worst correlated stems removed first). It shows that even with 10 stems, the SCT has very good reliability (.92), and that even at 5 stems the reliability is good (.85).

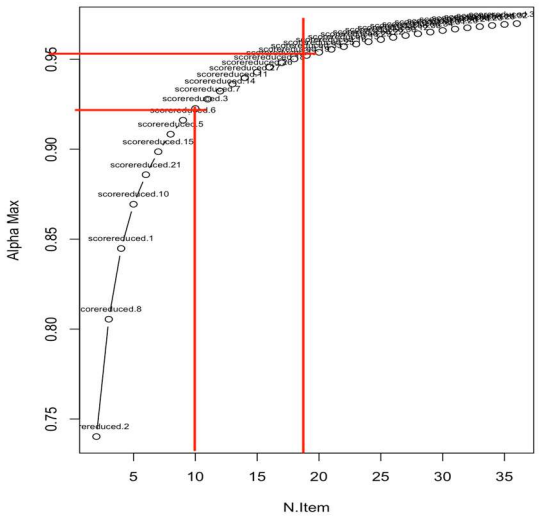


Figure 8: Crombach-Mesbah curve

Factor analysis. As concluded by prior research on the WSUCT, the SCT loads on one factor, i.e. seems to be measuring one latent variable. I.E. among the 36 stems none seem to cluster together as if they were measuring a sub-skill of the overall capacity. See Figure 9. "Factor/Component Analysis." (Note that, outside the scope of this report, even in analyzing the Specialty Protocols, the 6 specialty stems do not seem to cluster as a group compared with the other 30 'general' stems.)

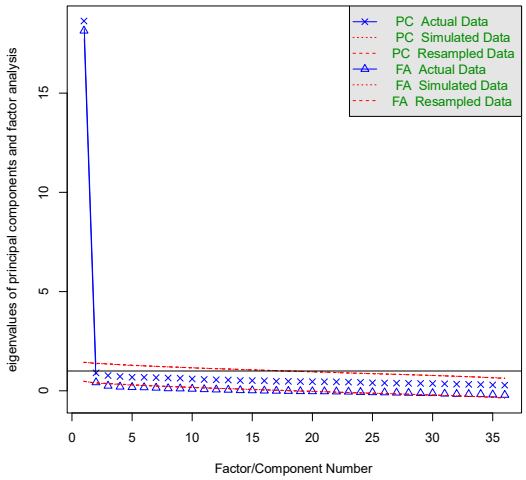


Figure 9. Factor/Component Analysis

Rasch analysis results

Though in our case the IRT analysis uses essentially the same (1-parameter) model as the Rasch analysis, we began the work using the IRT statistical packages in the R analysis language, along with our two IRT consultants, and later went into more depth using specialize Rasch analysis software (WINSTEPS) in consultation with our two Rasch consultants. For the WINSTEPS analysis we limited the data to only those 448 scores by the primary master Scorer (O'Fallon). This eliminated the possibility that results were influenced by variations in inter-rater reliability (as suggested by a consultant). Our analysis used the "Andrich's Rating Scale Model" (RSM) within the Rasch system (the simpler of two alternatives, as the data fit the Rasch model well).

WINSTEP Person and Item Distributions. Figure 10 shows the primary, or most basic, type of output from the WINSTEP (Rasch) analysis program. She bar chart on the left shows the distribution of subjects (surveys) and the bar chart on the right shows the distribution of items (stems)—both of which are in terms of the logit-scale "measure" normalized to go from -4 to +4. What our experts can tell from this chart is that:

- The "test coverage is impressive" over the range
- The distributions are relatively normal, which is good. (One can see the dent in the curve representing lower occurrence at the 4.0 level mentioned above.)
- The items are all clustered around a similar difficulty (the mean difficulty), as noted above. This is unusual for tests in general but was expected for our "projective test."

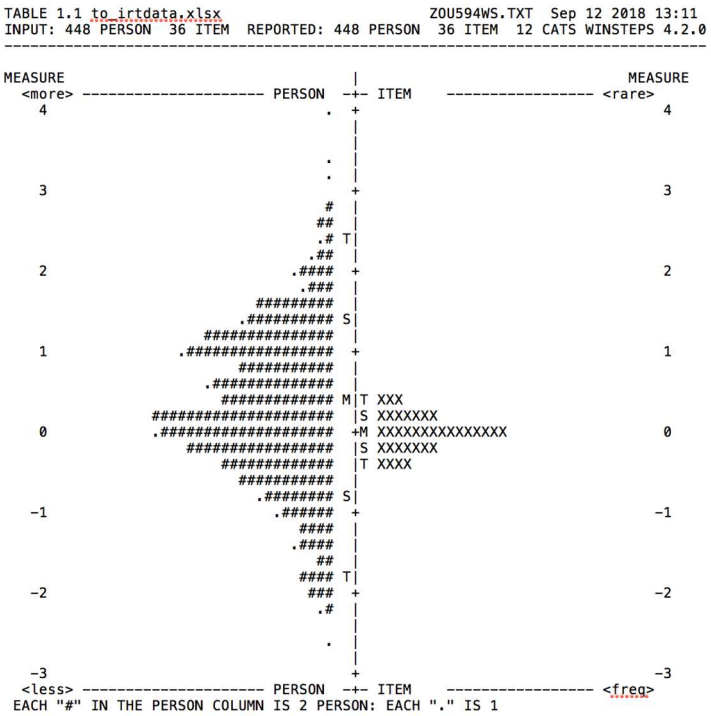


Figure 10: Person and Item Distributions from WINSTEP analysis

Other output tables in the WINSTEP analysis confirm what has been said above, that the test would be just as powerful with many less items. From a psychometric perspective it would be said that many items were "redundant," but for the STAGES assessment in its normal use, the full set of 36 items gives scorers and debriefers needed additional information for the client.

Probability Category Curve. Figure 11 shows the PCCurve summarizing all of the data. For example, the "3.5" curve shows the probability of scoring an item at any of the levels, for someone with an overall 3.5 average score. The first and last curves (1.0, 6.5) are always ignored in such plots (due to the math the go to infinity).

What is striking about the graph is that every other curve is (usually) lower. In looking at how/where the curves cross each other one who knows how to read such graphs will say that the scoring system is not good at distinguishing every other level. This is like saying that the scoring can strongly differentiate the 6 person perspectives, but is not very strong at differentiating early PP from late PP, i.e. the X.0 from the X.5. (Actually the analysis is ambivalent in concluding that *either* the X.0 and the next x.5 get mashed together; *OR* that the X.5 gets mashed together with the next X.0; but we will assume it is the former.) I.E. the scoring well discriminates 6 levels ("response categories") but not 12, at least for the data at hand.

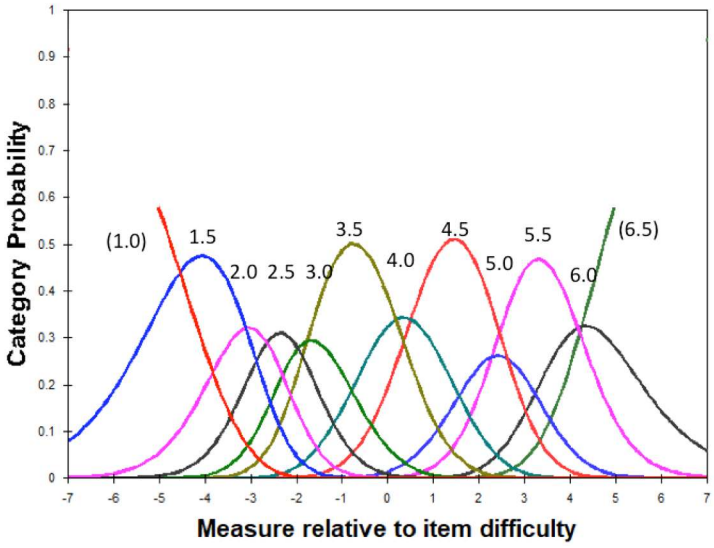


Figure 11: Probability Category Curve

This chart also shows that the separations between values is relatively even over the range, which is one of the primary predications by O'Fallon.

Summary and Fit table; and Strata. Figure 12 shows the Summary table with fit statistics from the WINSTEP output. INFIT and OUTFIT statistics assess outliers and possible problems. The fit statistics for misfitting items is "very good" (values less than 2 are considered acceptable). (It is not as good for the Persons as for Items, indicating the sample had a non-trivial number of people with non-standard response patterns.) The Reliability of the Items, .95, is excellent, and is analogous to the excellent Cronbachs alpha noted previously.

Of particular note is the Separation Index of Persons, 5.52. This, through a transformation formula, gives an indication of the number of "**strata**" or levels that the test can differentiate. (It is estimated as the adjusted or "true" item standard deviation divided by the average measurement error, expressed in standard error units.) The formula for converting the Person Separation Index to the number of discernable strata is (4S-1)/3. So (4*5.52-1)/3 = 7.0 strata. This is analogous to the estimate given above with the PPC curve that the test cleanly differentiates only about 6 levels (7 according to this calculation).

PERSON	448 INPUT	448 MEASURED			INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	286.3	36.0	.31	.18	.99	-.2	.99	-.2
P.SD	36.0	.0	1.03	.03	.50	1.7	.50	1.7
REAL RMSE	.18 TRUE SD	1.02	SEPARATION	5.52	PERSON RELIABILITY			.97
ITEM	36 INPUT	36 MEASURED			INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	3562.5	448.0	.00	.05	.99	-.2	.99	-.2
P.SD	96.8	.0	.22	.00	.16	2.2	.17	2.3
REAL RMSE	.05 TRUE SD	.22	SEPARATION	4.43	ITEM	RELIABILITY		.95

Figure 12: Winsteps Summary for Protocal 4 Rater #2

The clear differentiation of 6 (or 7) levels holds for our particular data set, which has some levels much more highly represented (see histograms in Appendix 2). Another metric given by WINSTEPS indicates the "maximum" strata number, which "reports how many statistically different ability levels could be distinguished in a *uniform* sample covering the full raw-score range of the test." This is independent of sample size, and is unlikely to be obtained for any sample (which tend to have a normal-shaped distribution). But it is pertinent to note that the "maximum strata" for our system is calculated to be 24 levels ("23.9); indicating that theoretically the system can differentiate at the level of the half-stage (suggesting that the strata index would improve with more data across all levels).

Step uniformity. Figure 13 is a graph showing the relationship between the STAGES score for each item vs. the Rasch analysis measure of the latent variable. The relationship is quite linear within the normal range of scores, and more importantly, the spacing between levels is relatively uniform. The latent variable represent the Rasch analysis estimate of true difficulty or skill level, and the score comes from the analysis of response text, which is intended to (indirectly) measure that latent variable. The relatively even spacing tentatively confirms O'Fallon's hypothesis that the STAGES model cuts the developmental space into equal-spaced units. O'Fallon has proposed that one benefit of a theory-based (i.e. explanatory vs. more descriptive and data-drive) model was that it offered a consistent "ruler" against which to measure the progression of stages.

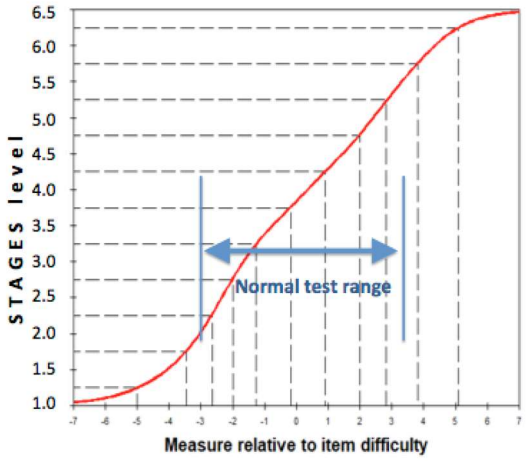


Figure 13: Stages score vs. Latent Variable Measurement

Focal Score—a new aggregation method

The IRT and Rasch analysis and consultants (1) gave evidence that the Ogive cutoff system was problematic, and (2) that the test is valid and strong. But this work did not easily yield a method to replace the existing Ogive method however. IRT (and

Rasch) analysis assumes that a sum or average is a sufficient score, and this would correspond roughly to the existing TWS metric. However, we wanted a metric that better captured the principle, well captured by the ogive method, that with a projective test: (1) lower scores are "throw-aways" because the subject is not "trying" to do their best, and can they include spurious or flippant responses (in fact higher developmental levels are assumed to be more playful in this way); and (2) the "center of gravity" or "total" score should prioritize or weigh the higher scores more, as they show the true competence of the subject (i.e. one is unlikely to randomly "guess" or fake very many high scores in the SCT).

We had two primary goals for the new aggregate method, which we call the Focal Score. First, to avoid all the problems of the ogive cutoff method and use something more straightforward, and second to minimize the differences between the new and old scores, so that we would not have to alter the established interpretation of each level (e.g. what a "3.5" total score means).

We considered two methods. The first was taking the average of the top X scores, where X was to be determined experimentally. The second method was a true weighted score, in which scores of higher levels received additional weights. Our method of assigning weights was to raise scores to a power P, where the power was to be determined experimentally. By "experimentally" we mean this: iterating over different values to see which one comes closest to the old ogive score. We has confidence that these methods would not have unusual statistical properties because (1) the IRT analysis showed the test was strong, and (2) the analysis of skew, kurtosis, and standard deviation across, for surveys at all levels, looked good.

The new aggregation methods we tried were:

- Average of the: top 5, top 8, top 10, top 12, and top 18 scores; and
- Sum of the scores raised to the power of: 1, 5, 8, 10, 12, 15. (Remember that scores are within 1.0, 1.5, ...6.5.)

Results are in Table 1. We compare each new method with the old ogive score, taking the root mean square error over the entire database (n=740). "Ave of all" is comparable with the TWS method (which is essentially the sum of the scores), it shows a baseline RME error of 0.55. Note that because the scores increase by 0.5, and RMSE of 0.5 is one full level. The lowest RMSE for the TopX method was 0.22 at Top10 and Top12. In taking the average of the sum of the Score raised to powers, the power with the lowest RMSE was Score^8. It appeared that this method did not produce errors as low as the TopX method did. In the final entry, we tried combining these two methods, and note that it did not improve the RMSE.

Table 1: RMS Error of Aggregate Methods

Method	RMSE

ave of all (like TWS)	0.55
aveTop5	0.30
aveTop8	0.23
aveTop10	0.22*
aveTop12	0.22
aveTop15	0.23
aveTop18	0.25
Score^1	0.48
Score^5	0.32
Score^8	0.26*
Score^10	0.27
Score^12	0.28
Score^15	0.33
max10score^2	0.24

The RMS error is an average error over all surveys. But we wanted to look closer how the methods performed at each level. It is more important that the methods work well where there are the most scores, *and* having low RMSE for higher scores was more important than for lower scores (Concrete tier) because we realized that our new method was re-conceptualizing the aggregate score at low levels. As discussed above, the ogive score conflates shadow crashes with meaning making complexity, and we wanted the new measure to focus on meaning making complexity (or maturity).

Figure 14 shows histograms showing the scores of the new methods (Y axis) with the old ogive score (x axis). The bar marked "SAME" is there for comparison to illustrate the "straight line" one would get if a new method gave the exact same score as the ogive method. We decided to use either Top6 or Top12, because they represent exactly 1/6th or 1/3rd of the 36 stems (the difference in the RMSE between Top6 and the lowest RMSE of Top10 was small). The chart shows how the error varies by center of gravity. Top6always performs better than Top12, and both always do better than the Average, so we choose (average of the) Top6 as the new Focal Score method.

Next, to get the Focal Scores even closer to the old ogive scores, we did a linear transformation of the Top6 score. Linear fitting showed that a formula of (1.2*Ave-top6)-1 would yield a score closer to the old ogive. We call this the "adjusted" Top6, and it is also shown in the chart. The adjusted top6 performs better for most of the subtle and metaware levels, and the fact that it does not do as well for the Concrete tier is, as mentioned, not important.

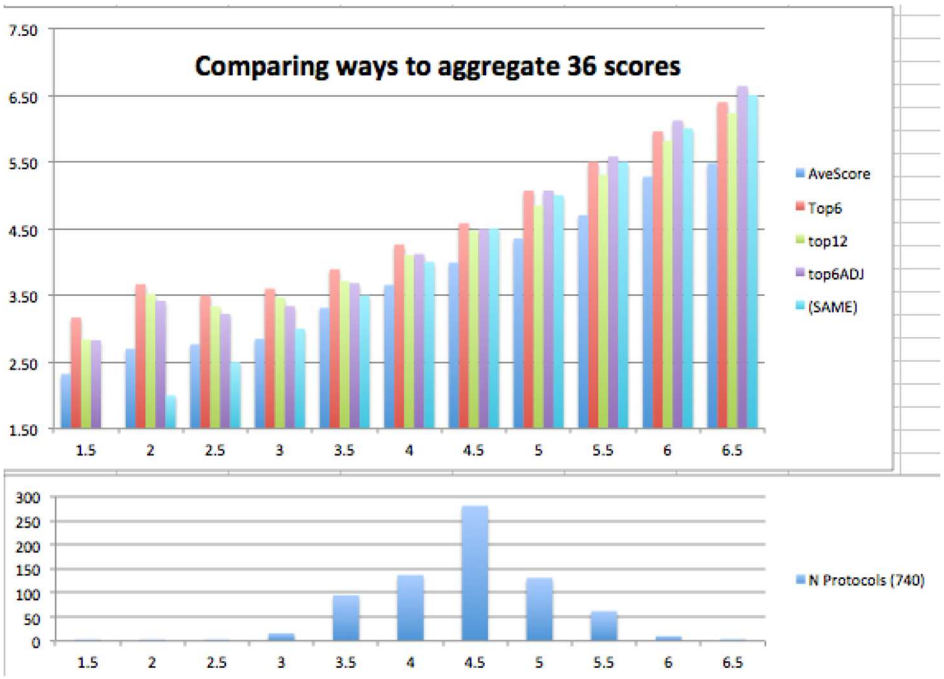


Figure 14: comparing aggregation methods across all levels

Focal Score. In sum, the new aggregate method, called the Focal Score, uses the formula **(1.2*Ave-top6)-1**. For protocols of arbitrary length, the Focal score takes the top 1/6th of the scores. Unlike the ogive score, which yields a category, the new Focal Score value yields a decimal number (which is usually rounded to one decimal digit).

- It yields results that are usually quite close to the old ogive score (but only on average).
- It is easy to calculate (uses a formula rather than a procedure).
- It behaves more predictably, like standard average or total scores used in psychometric analyses.
- It does not confuse shadow with complexity maturity, and thus will be useful for scoring children.
- It does not contain a large number of arbitrary parameters (such as the ogive cutoffs and the TWS multipliers).
- It is not based on suspicious application of Bayes theorem.
- It does not force scores into categories but lets them have fractional values.
- It is not sensitive to small differences in how stems are scored (which happens in the ogive method near the cutoffs).
- Also, the new Focal Score rating makes it easier to produce a total score for protocols of any length (one does not have to worry about inventing new cutoffs). I.E. one takes the average of top 1/6th of the scores (this works easiest for protocols having a multiple of 6 stems).

This Focal Score measure has been added to the STAGES Training Platform, and will seen be included in the Scoring Platform soon. The old ogive and TWS values will always be shown (perhaps off in a corner) for comparison. For several months Terri has been observing on the Training Platform that the Focal Scores match with her intuitions about stages very well, close to the old Ogive score, and work even better than the old Ogive score in the Concrete tier.

Additional measurements: Average, Bottom, Spread. The new scoring system shows several measures in addition to the Focal Score: The *Average* (across all stems); *Bottom Score* (the average of the bottom 6, related to a "shadow score" for adults); and *Spread* — the spread score is the Focal Score minus the Bottom Score, which is an indication of the range of values in the survey.

Early and Late stages. Now that the total score allows for fractional values, there are two ways to think about early and late (and middle) scores for any stage. We will need to think more about how these two methods will coordinate with each other. The existing (old) method uses clues in the response language to amend a total score as "early" or "late." With the new Focal Score, early, mid, and late could mean simply where the fractional score is along the continuum. These two methods will not necessarily coincide! Terri will keep an eye on this question and we will discuss it later.

Also, the question arises RE where along the fractional scale we call "early" and "late." For example, early 3rd person perspective could be just below "3.0" as in "2.9 or 3.0", or it could be just above 3.0, as in "3.0 and 3.1." After consultations with Terri we decided tentatively to use the convention shown in Table 2, and illustrated in Figure 15.

Table 2: Decimal fractions vs. early and late stages

X.0 early	0.5 early
X.1 mid	0.6 mid
X.2 mid	0.7 mid
X.3 late	0.8 late
X.4 transition between	0.9 transition between

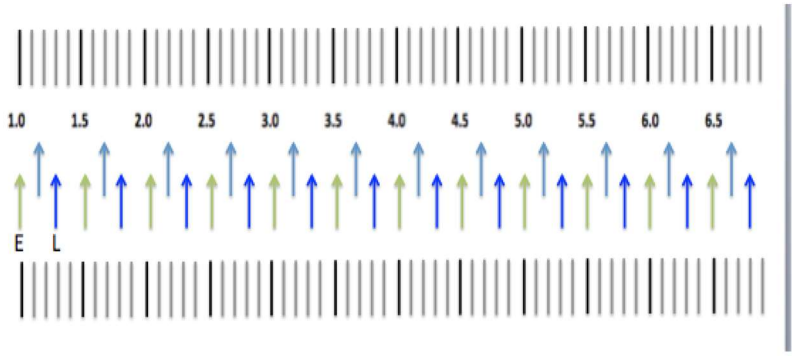


Figure 15: Early and Late stages, illustrated.
(Early in green, late in dark blue, mid in light blue)

Also, we may want to know what to add to scores if the text indicators indicate early or late stages. The Focal Score could be adjusted in the following way if early or late indicators were found:

- for Early add 0
- for unspecified or middle, add 0.15
- for Late, add 0.35

Conclusions and Future work

This research has increased our confidence in the STAGES theory, model, and scoring methodology. It has also produced a new method for calculating the aggregate (total) score over the 36 stems (or fewer), which we call the Focal Score. The Focal Score has many desirable properties, as notes in the prior section. The STAGES Platform will also showing the additional metrics: Average, Bottom score, and Spread.

Future research. Here are some things we could explore further:

- RE the "strata" metric, and poor ability to differentiate all 12 levels: it might be possible to maodify the coding manual to better differentiate early from late stages (and active vs passive). One method for doing this would be to compare surveys with Focal Scores around "X.5" vs those at X.0 above and below them; and look for additional principles that would differentiate the X.5's.
- We could look more closely at the differences between stem items: their average difficulty and standard deviations; and which levels they do a better job at differentiating (see the "information curve" Figure).
- We could look more closely at what is going on with the "I am" stems; and see if additional stems could be invented that target the same sub-capacity that it does.
- We could investigate dip in the number of stems scored at 4.0 over all items
- Given the predominance of 4.5 surveys, it would be good to get more data overall in the database for 2.5, 3.0, and 3.5 stages.

References

Andrich, D. (1988). *Rasch models for measurement* (Vol. 68). Sage.

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Hillsdale, NJ: Erlbaum.

Commons, M. L., & Pekker, A. (2009). Hierarchical complexity and task difficulty. *Unpublished manuscript. Available at http://dareassociation. org/papers*

Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A., & Krause, S. R. (1998). Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review*, 18(3), 237-278.

Cook-Greuter, S. R. (1999). *Postautonomous ego development: A study of its nature and measurement* (Doctoral dissertation, Harvard Graduate School of Education).

Cook-Greuter, S. R. (2002). A detailed description of the development of nine action logics adapted from ego development theory for the leadership development framework. *Retrieved Oct*, *13*, 2002.

Cook-Greuter, S. R. (2004). Making the case for a developmental perspective. *Industrial and Commercial Training*, 36(7), 275-281.

Dawson-Tunik, T. L., Commons, M., Wilson, M., & Fischer, K. W. (2005). The shape of development. *European Journal of Developmental Psychology*, *2*(2), 163-195.

Dawson, T. (2004). Assessing intellectual development: Three approaches, one sequence. *Journal of adult development*, 11(2), 71-85.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*, 477-531.

Holt, R. R. (1980). Loevinger's measure of ego development: Reliability and national norms for male and female short forms. *Journal of Personality and Social Psychology*, *39*(5), 909.

Loevinger, J. (1979). Construct validity of the sentence completion test of ego development. *Applied Psychological Measurement*, *3*(3), 281-311.

Loevinger, J. (1985). Revision of the sentence completion test for ego development. *Journal of personality and social psychology*, 48(2), 420.

Loevinger, J. (Ed.). (1998). Technical foundations for measuring ego development: The Washington University sentence completion test. Psychology Press.

Loevinger, J., & Wessler, R. (1970). *Measuring ego development: Construction and use o fa Sentence Completion Test, Vol. 1.* San Francisco: Jossey-Bass.

Murray, T. (2017). Sentence completion assessments for ego development, meaning-making, and wisdom maturity, including STAGES. *Integral Leadership Review*, August, 2017.

O'Fallon, T., Polister, N., Blazej Neradiek, M., Murray, T. (in preparation). STAGES: A New Integral Scoring Methodology for Perspectival Levels in Ego Development.

O'Fallon, T. (2011). STAGES: Growing up is Waking up—Interpenetrating Quadrants, States and Structures. Available at www.pacificintegral.com.

O'Fallon, T. (2012). Development and consciousness: Growing up is waking up. *Spanda Journal*, III(1), 97-103.

O'Fallon, T. (2013, July). The senses: Demystifying awakening. Presented at the Integral Theory Conference.

O'Fallon, T. (June 2010a). Developmental experiments in individual and collective movement to second tier. *Journal of Integral Theory and Practice,* 5(2), 149-160.

Rasch, G. 1980. *Probabilistic model for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Torbert, W. R. (2014). Brief comparison of five developmental measures: The GLP, the MAP, the LDP, the SOI, and the WUSCT primarily in terms of pragmatic and transformational validity and efficacy. Available from Action Inquiry Associates, www.williamrtorbert.com.

Torbert, W. R., & Livne-Tarandach, R. (2009). Reliability and validity tests of the Harthill Leadership Development Profile in the context of Developmental Action Inquiry theory, practice and method. *Integral Review*, *5*(2), 133-151.

Appendices

Appendix 1: TWS calculation and Ogive Cutoffs

The original STAGES TWS formula, an extension of the one used by Cook-Greuter and Loevinger, multiplies the frequencies of each level by a factor, and sums those results, using these factors:

LEVEL	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5
FACTOR	2	3	4	5	6	7	8	9	10	11	12	13

The Table below is used to determine the final derived score. The STAGES model uses the identical approach utilized by Loevinger (1998) and updated by Cook-Greuter (1999) but adds three levels (2.0, 5.5 6.5).

Loevinger and Cook-Greuter use a "cutoff" method they called the "Ogive" method to aggregate the scores of the 36 completions to produce an overall "center of gravity" score for each inventory. The cutoff method is a procedure where each step says "if there are X or more completions at level L then the final score is L"—where X differs depending on the level. Below is a summary of the cutoff approach used for the STAGES model.

Note how the levels being tested start with the highest (6.5), move down to 3.0, and then start at the lowest and work up to 2.5. This is because the cutoff numbers were determined by CG/L using a method that they linked to Bayes Theorem. The 2.5 (Diplomat) level was estimated by Loevinger (and Cook-Greuter) to be the most prominent in the normal population and thus the most likely given no additional evidence. This is why 2.5 is the final or "default" value in the procedure.[3]
Cook-Greuter used the cutoff of four for the two levels above Strategist (4.5), and STAGES duplicated that approach, using 4 for all stages above 4.5.

Table 4: Cutoff Rules For a 36–Item Sentence Completion Test*

(Do steps in order and stop where true) **If there are:**			**Assigned stage**
4	or more rated at	6.5 or higher	6.5
4	or more rated at	6.0 or higher	6.0
4	or more rated at	5.5 or higher	5.5
4	or more rated at	5.0 or higher	5.0
6	or more rated at	4.5 or higher	4.5

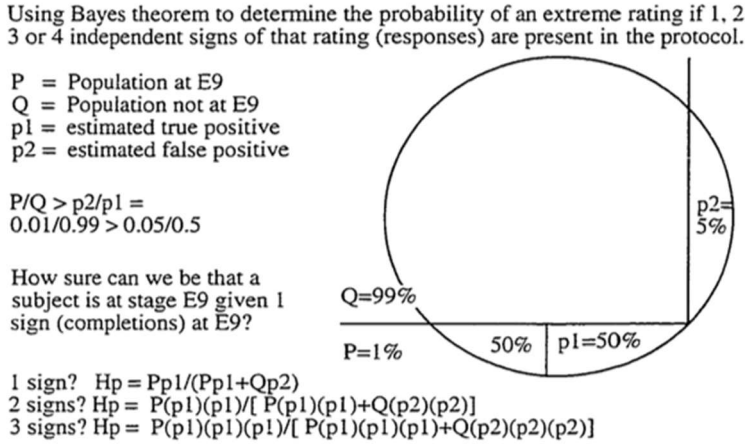
[3] This Ogive method contains a number of assumptions and approximations which we will discuss in a future paper. What is important here is that the method is the de-facto standard that has been used in the numerous research studies using ego development assessment for almost 40 years.

9	or more rated at	4.0 or higher	4.0
14	or more rated at	3.5 or higher	3.5
17	or more rated at	3.0 or higher	3.0
7	or more rated at	1.0	1.0
7	or more rated at	1.5 or lower	1.5
7	or more rated at	2.0 or lower	2.0
-	If none of the above, use		2.5

** Starting at the highest stage, 6.5, and proceeding down the table, assign the stage from the first row where the rule applies.*

The cut point criterion in the higher stages 5.0, 5.5, 6.0, 6.5 is analogous to the cut points in the Cook-Greuter scale.

Below is a summary from Cook-Greuter's dissertation (1999) explaining how cutoffs were calculated using Bayes theorem, followed by an explanation from Loevinger (1998). Loevinger says "The rule is not helpful at extreme ratings. However, the number of cases per category at the extremes is so small that theory or intuition is a more reliable guide" (pg. 121) .



Cook-Greuter 1999, Fig 4.11 A

Calculations if P=1%, Q=99%, p1=50%; p2=5%.
How sure are we with 1 sign that a protocol is at the rare stage C9?
$Hp = Pp1 \div (Pp1+Qp2) = (0.01)(0.5) \div [(0,01)(0.5)+(0.99)(0.05)] = 0.005 \div (0.005+0.0495)$
$= 0.025 \div 0.0545 = 0.09 \rightarrow 10\%$
How sure are we with 2 signs? $Hp = P(p1)^2 \div [P(p1)^2+Q(p2)^2]$
$= (0.01)(0.5)(0.5) \div [(0.01)(0.5)(0.5)+(0.99)(0.05)(0.05)] = 0.50 \rightarrow 50\%$
How sure are we with 3 signs? $Hp = P(p1)^3 \div [P(p1)^3+Q(p2)^3] = 0.91 \rightarrow 91\%$
How sure are we with 4 signs? $\rightarrow$ 99%;

Loevinger's calculations if P=5%, Q=95% then
1 sign p $\rightarrow$ 34%; 2 signs $\rightarrow$ 84%; 3 signs $\rightarrow$ 98%.

Cook-Greuter 1999 Fig 4.11 B

A category *i* is located at level *Ek* when both of the following conditions are met:

1. The probability that a person who has a TPR at or below level *Ek* will give a response in category *i* exceeds the probability that a person with TPR above level *Ek* will give a response in category *i*.
2. The probability that a person who has a TPR at or above *Ek* gives a response in category *i* exceeds the probability that a person with TPR below *Ek* will do so.

The probability that a person with given TPR *Ej* will give a response in category *i* is n_{ij}/N_j, where N_j is the number in the sample at level *j*, and n_{ij} is the number in category *i* at level j, that is, with a TPR of *Ej*. (*Ej* goes from E2 to E9.)

$$1.\ \sum_{2}^{k} n_{ij}/N_j > \sum_{k+1}^{9} n_{ij}/Nj$$

$$2.\ \sum_{2}^{k-1} n_{ij}/N_j < \sum_{k}^{9} n_{ij}/Nj$$

Loevinger, Technical Foundations..., App. E

Appendix 2: STAGES data descriptive statistics and charts

Some notes on the charts below:

Figure A2. 1: Histograms of scores per level: A: per response; B: per protocol

- Overall there are many more protocols in the database at 4.5, showing a skew toward that level in STAGES scored clients.
- Note the curious dip at 4.0 for response scores. Its a bit of a mystery why there are so fewer 4.0 scores.

Figure A2.2: Comparing protocol scores for STAGES vs SCTi (MAP) scores in the database

- This Figure just illustrates that the older data in the STAGES platform database, which was scored using the older MAP method, has a different spread of scores vs the more current STAGES scores. It points to the potential to compare STAGES vs MAP data in future research questions.

Figure A2.3: Histograms of responses for each stem

- This Figure illustrates that there may be interesting differences in how each stem elicits scores from different levels. Suggesting the possibility for further analysis at the stem level.
- Of particular interest are patters that seem to alternate every other level, corresponding to active vs. passive biases.
- This graph is essentially additional detail on Figure 6, showing "High and Low Averages and Standard Deviations for Stems."

Figure A2.4: Per-stem histograms, with normalized version of the left side.

- This Figure shows per-stem histograms as in the prior Figure, but add normalized versions that highlight how each stem differs from other stems at each STAGES level. Again, Suggesting the possibility for further analysis at the stem level.

Figure A2.5: Item correlations.

- This Figure shows a colored "heat map" of the correlation between every item with every other item. Items are reasonably well correlated, with correlations ranging from .32 to .53.

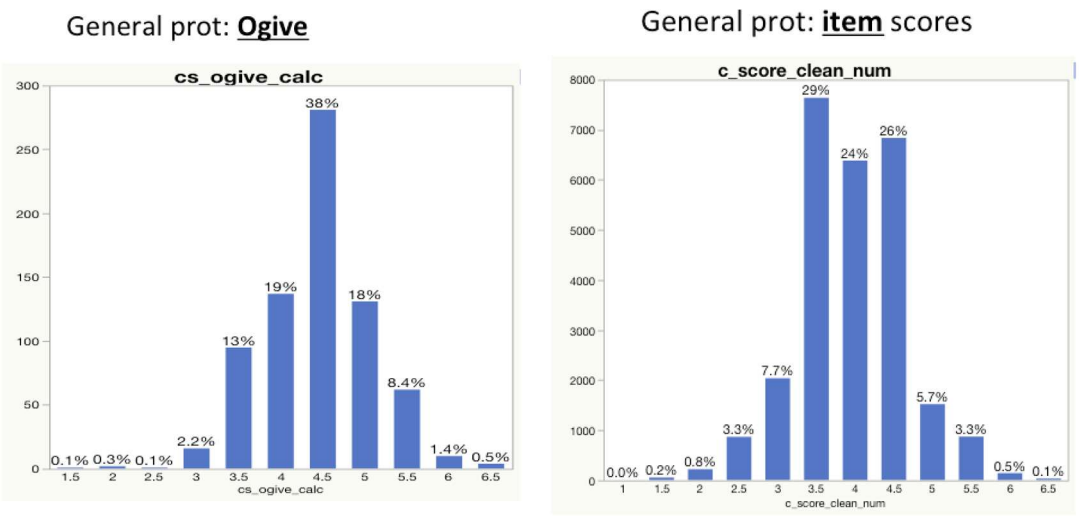


Figure A2. 1: Histograms of scores per level: A: per protocol; B: per response

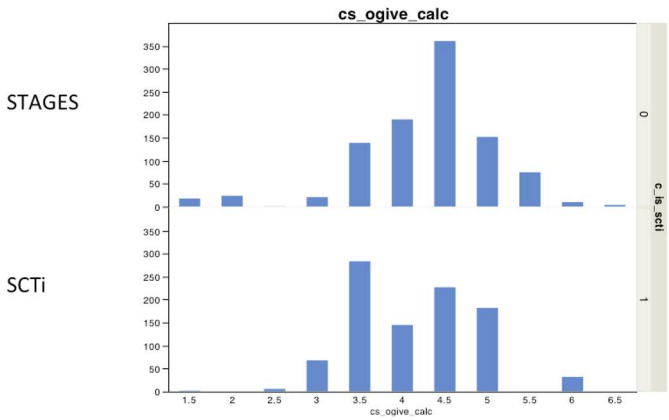


Figure A2.2: Comparing protocol scores for STAGES vs SCTi (MAP) scores in the database

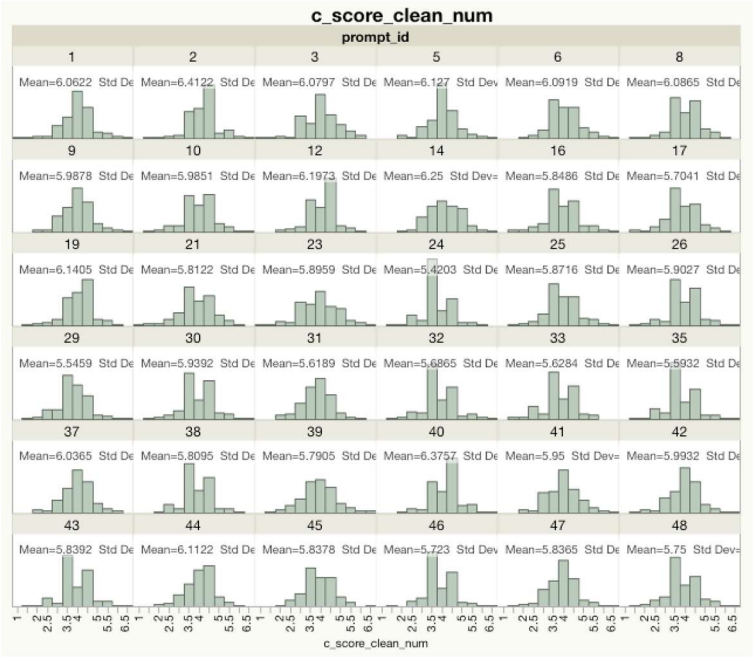


Figure A2.3: Histograms of responses for each stem

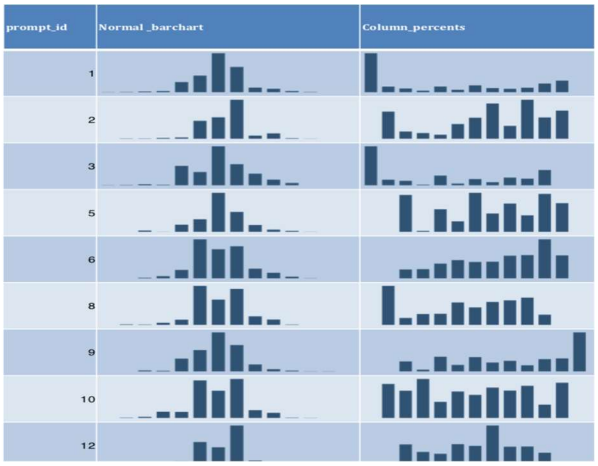


Figure A2.4: Per-stem histograms, with normalized version of the left side.

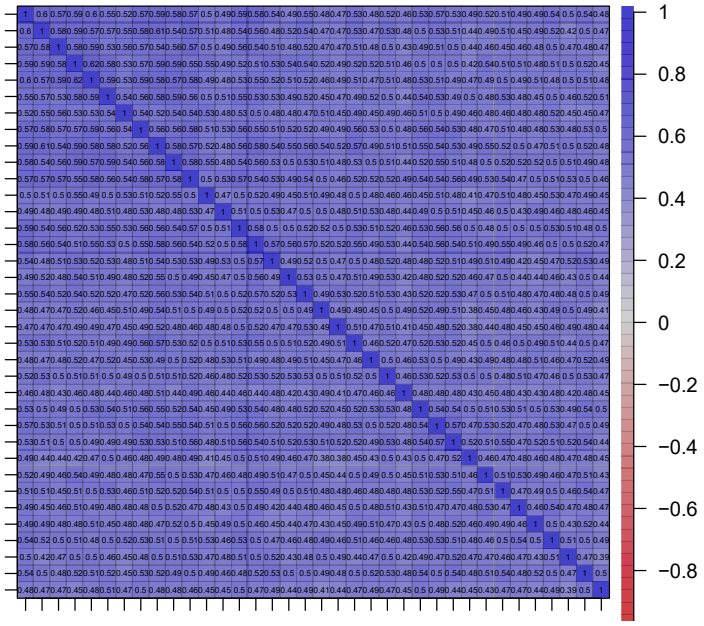
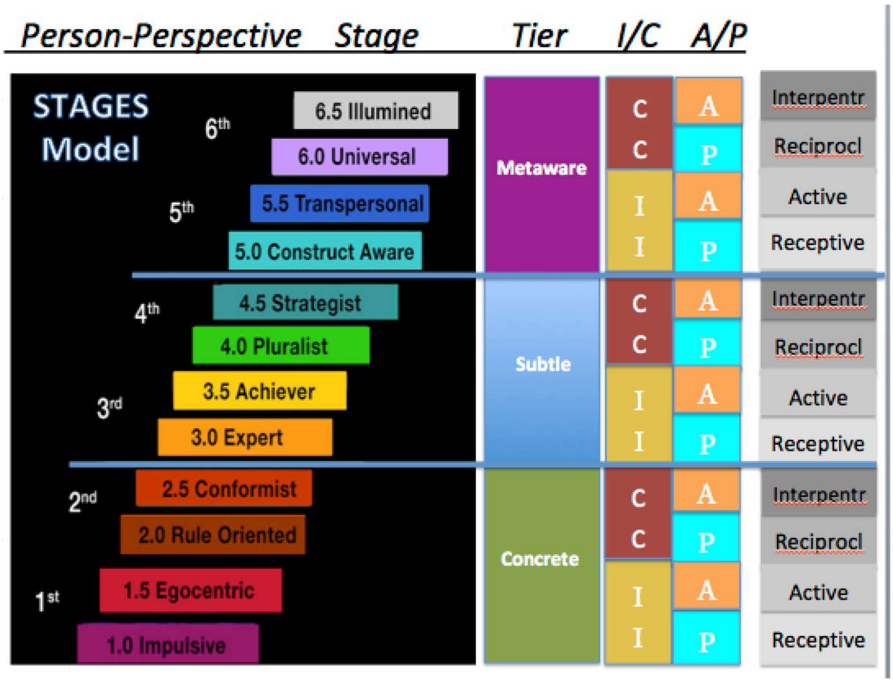


Figure A2.5: Item correlations

Appendix 3: Stages Model Figure



Appendix 4: IRT vs Classical test theory

This table illustrates key differences between classical test theory and Item Reonse Theory (including Rasch analysis).

CTT	IRT
▪ The test is the unit of analysis	▪ The item is the unit of analysis
▪ Measures with more items (longer) are more reliable than their counterparts	▪ Measures with lesser items (shorter) can be more reliable than their counterparts
▪ Comparing scores from different measures can only be done when the test forms/ measures are parallel	▪ Item responses of different measures can be compared as long as they are measuring the same latent trait
▪ Item properties depend on a representative sample	▪ Item properties don't depend on a representative sample
▪ Position on the latent trait continuum is derived from comparing the test score with scores of the reference group	▪ Position on the latent trait continuum are derived by comparing the distance between items on the ability scale
▪ All items on the measure must have the same response categories	▪ Items on a measure can have different response categories

end