

# Investigating the Validity of the Ogive Aggregation Method, Including the use of Rasch Analysis, for the Sentence Completion Test and the STAGES Model

Tom Murray<sup>1</sup>

**Abstract:** This project assesses psychometric aspects of the STAGES sentence completion test (SCT) using data from 740 scored surveys, and some of the analysis applies to all variations of the SCT for ego development (meaning making maturity). The goals of this research project include: (1) to apply item response theory (IRT) and Rasch analysis to determine item-level psychometric properties of the SCT that were previously unaddressed in SCT research; and (2) to further investigate suspected problems with the ogive cutoff method for aggregating item scores in the SCT and propose alternatives. The psychometric analysis includes: within-test item normality, item standard deviations, test length analysis, factor analysis, characteristics of and correlations among each item, overall test strength, and construct levels discrimination. A range of issues with the standard ogive cutoff method are described, and a new item aggregation method is then proposed.

**Keywords:** Adult ego development, ogive cutoff, psychometrics, STAGES, sentence completion test.

## Executive Summary

This project assesses certain psychometric aspects of the STAGES sentence completion test (SCT) using data from 740 scored surveys, and some of the analysis applies to all variations of the SCT for ego development (meaning making maturity). The goals of this research project include: (1) to apply item response theory (IRT) and Rasch analysis (a subset of IRT) to determine item-level psychometric properties of the SCT that were previously unaddressed in SCT research; and (2) to further investigate suspected problems with the ogive cutoff method for aggregating item scores in the SCT and propose alternatives.

---

<sup>1</sup> **Tom Murray** is Chief Visionary and Instigator at Open Way Solutions LLC, which merges technology with integral developmental theory, and is also a Senior Research Fellow at the University of Massachusetts School of Computer Science. He is an Associate Editor at Integral Review, is on the editorial review board of the International Journal of Artificial Intelligence in Education, and has published articles on developmental theory and integral theory as they relate to wisdom skills, education, deliberative skills, contemplative dialog, leadership, ethics, knowledge building communities, epistemology, and post-metaphysics. See [www.tommurray.us](http://www.tommurray.us) [tommurray.us@gmail.com](mailto:tommurray.us@gmail.com)



The paper begins with an overview of the history of the SCT, including Loevinger's original work and variations by Cook-Greuter, Torbert, and O'Fallon. We then describe the STAGES model and assessment in greater detail, highlighting how the model differs from prior SCTs.

We summarize evidence for the psychometric strength of the SCT. Most of the 400+ studies of the SCT were conducted with Loevinger's WUSCT, but studies of the other variations inevitably replicate results of the general validity characteristics of the test. We summarize prior research on the SCT regarding inter-rater reliability, internal consistency, test-retest reliability, face validity, construct validity, incremental validity, clinical utility, and predictive validity, supporting Westenberg et al.'s conclusion that "findings of over 350 empirical studies generally support critical assumptions underlying the ego development construct" (p. 485).

Specific to the STAGES model, we reference another peer-reviewed paper (O'Fallon et al., 2020) describing a "replication study," which shows that the STAGES scoring method replicates prior (MAP) SCT scoring up to 4.5 Strategist. That paper also summarizes longitudinal data analysis including the upper levels, which shows evidence for the monotonic developmental sequencing of O'Fallon's newly defined highest stages (Metaware Tier).

We next focus on methods used to aggregate the 36 item scores into the total protocol rating (TPR), especially the ogive cutoff method and secondarily the less-used TWS (total weighted score) method. We argue that the ogive method entails a number of drawbacks, including:

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

This paper then describes the method and results of our statistical analysis of the 740 STAGES protocols using descriptive statistics, item response theory, and Rasch analysis. Below is a summary of the findings, which hold for levels 2.0 to 6.0, however for the rare extremes of 1.0, 1.5, and 6.5 surveys, there are too few data points for confident conclusions.

1. **Within-test item normality.** A skew and kurtosis analysis shows that scores within each survey are, on average, normally distributed. This does not support the common supposition that the scores' "bell curve" within surveys would contain more items below the average score than above it.
2. **Item standard deviations.** The average standard deviation of item scores within a survey is remarkably similar across all developmental levels. Thus, the shape of the distribution of the 36 scores within a survey does not change much for individuals at different centers of gravity.

3. **Test length.** A Cronbach-Mesbah analysis (of the Cronbach's alpha score for successively fewer test stems) shows that for a half-item test the reliability is excellent (.95); that with 10 stems, the SCT has very good reliability (.92); and that at 5 stems the reliability is acceptable (.85). This supports any future projects that use surveys with fewer than the 36 items.
4. **Factor analysis.** A factor/component analysis of the data confirms prior findings that the SCT loads on one factor or in other words seems to be measuring a single latent variable.
5. **Characteristics of and correlations among each item.** We show measures of difficulty (mean score) and spread (standard deviation) for each test item to test the assumption that the SCT test items designed to triangulate (parallax) the construct from many life-contexts are generally equivalent (though they will vary in how they function per individual). The results confirmed in the Rasch analysis show that there is some variation among the items, but overall they have similar difficulty and spread. However, the stem "I am..." stands out as being the least highly correlated with all other stems and as having the highest standard deviation. We also show a heat map illustrating the 36x36 item correlation magnitudes.
6. **Overall test strength.** Rasch and IRT analyses confirm that the test is very robust/sound as a psychometric measurement across its entire range. Test reliability and coverage are excellent and the items have good discrimination. Scoring is highly consistent (i.e., good quality) across the items (similar to prior Cronbach's alpha results).
7. **Construct levels discrimination.** The test easily differentiates the 6 person perspectives and, consistent with the model's hierarchical organization of 3 parameters, has more difficulty differentiating between levels at the granularity of 12 levels. The 12 stages defined by the model are relatively evenly spaced along the spectrum, which supports O'Fallon's conjecture that the STAGES model cuts the range into equal slices or provides a consistent "ruler" which is resilient to data collected in future studies from different populations.

**New item aggregation method.** Given the problems described for the ogive cutoff method, we experimented with a number of alternative methods for aggregating the survey items. Our goals for a new method were:

1. achieve compatibility with the expectation that for a projective test, a person's true "center of gravity" will be evidenced from the higher scores;
2. avoid the problems of the ogive cutoff method and use something more straightforward and with fewer arbitrary parameters;
3. to minimize the differences between the new method and the prior ogive method, so that we would not have to alter the established interpretation of each level (i.e., to remain consistent with how a "3.5/Achiever" total score is understood in the field).

The methods we tested are described in the paper. In the end we chose taking the mean of the top 6 scores in the survey (or the top 1/6th of the scores if the survey has other than 36 items). This is modified slightly to produce a value that is, on average, close to the original ogive method by using the formula:  $(1.2 * \text{AveOfTop6}) - 1$ . We call this value the Core Stage (for continuity with prior articles, note that we originally called it the Focal Score). STAGES International will begin to phase in this method, firstly alongside the ogive method, before eventually replacing it over the next year or two.

The Core Stage regards a continuous value. The benefits of this method include that the value is not sensitive to small perturbations or errors around cutoff boundaries and that it is easily extended to different test lengths without revising the set of cutoffs. Another benefit of this method is that it does not confuse shadow crashes with low complexity/maturity, which is the case with the ogive method, and thus can be used for scoring children (which O'Fallon has begun doing).

The revised scoring system will show several measures in addition to the Core Stage: the average (across all stems); the bottom score (the average of the bottom 6 related to a "shadow score" for adults); and *Spread* – the spread score is the Core Stage minus the bottom score, which indicates the range of values in the survey.

The final section of the report discusses future work and limitations, which largely pertain to available data.

The STAGES framework will be gradually shifted to the new Core Stage method. We remain agnostic regarding whether other SCT frameworks should adopt this method or any alternative to the ogive method, which has been performing well enough over the years and represents the default standard, and the effort to adopt a new method may well outweigh any benefits. The benefits of changing are greater for the STAGES model vs. other SCT variations, because others use a fixed set of stems while STAGES allows creating valid surveys with new stems of any length – which would require ongoing adjustment of cutoff values that can only be performed in arbitrary ways. In addition, automated computer scoring exists for the STAGES model (see [www.stagelens.com](http://www.stagelens.com)), which means the SCT will become more practical to use for large-scale research where the validity of methods is more critical.

## Introduction and Background

### Motivations and Research Goals

The sentence completion test (SCT) for adult development, originally created by Jane Loevinger and later extended by Susanne Cook-Greuter, William Torbert, and Terri O'Fallon (each developing different variations), is arguably the most thoroughly researched, validated, and used developmental instrument in adult psychology. Loevinger used the term "ego development" for this "holistic" view of personality and cognition that "[sees] behaviors in terms of meaning or

purposes" (Loevinger 1970, p. 3).<sup>2</sup> Ego development has been described using concepts including leadership maturity, perspective-taking complexity, sophistication of world-view consciousness, and "wisdom skills" and has substantial overlap with the construct of meaning-making maturity described in Kegan's construct developmental theory (Kegan, 1998). Kegan describes it as a "consistency in the structure, or order of complexity, of one's meaning-making (i.e. how one thinks)" (1998, p. 55) about the relationships between self, other, and the world (intrapersonal, interpersonal, and cognitive in Kegan's terms; and "I/we/it" in terms of Wilber's Integral Theory (1995)).

Loevinger's theory of ego development was intricately linked to her assessment instrument called the Washington University Sentence Completion Test (WUSCT; Hy & Loevinger, 1989; Loevinger & Wessler, 1970). The literature on the SCT is extensive, and according to an overview by Westenberg et al. (2004), the SCT test has robust psychometric properties with "indicated excellent reliability, construct validity, and clinical utility" (p. 596). They further state that, "findings of over 350 empirical studies generally support critical assumptions underlying the ego development construct" (p. 485), and dozens more studies have followed since 2004 (Torbert, 2009).

The SCT is a "projective" test in which subjects complete sentence starters and respond freely without a need to produce a "correct" or superior answer – which the theory claims affords the test analyst a deeper view into tacit or unconscious psychological traits. The standard SCT has 36 sentence starters ("stems") such as "Raising a family..." and "When people are helpless..." which the test taker completes (e.g., "...is a joy", "...they get taken advantage of") and are rated independently. Sentence completions vary from a few words (even one word) to full paragraphs (and rarely multiple paragraphs). Rather than a simple sum or mean of the item scores, the TPR score uses a more complex cutoff method called the "ogive method" for reasons we describe later. This converts a complex continuous score into one of a set of discrete categories (levels or stages: 8 for Loevinger's model, 9 for Cook-Greuter's and Torbert's models, and 12 for O'Fallon's model).

Despite the depth and breadth of the SCT instrument and its psychometric strength, assessments of its validity all (as far as we can tell) have used "classical test theory" and focused on the validity of the assessment as a whole (i.e., the TPR). There has been little focus on its per-item psychometric properties besides the internal consistency of the items (the Cronbach's alpha measure). Although it is still widely used due to its robustness in most contexts, classical test theory has been superseded in the last half century by item response theory (IRT) psychometric methods, which include a deeper analysis at the item level of the test. The present study applies IRT (and Rasch analysis) to SCT data to gain insights about the properties of the SCT. More importantly, the ogive method for aggregating item scores has been largely taken for granted and simply re-used in the hundreds of studies using the SCT, despite some limitations of the ogive method.

---

<sup>2</sup> Browning (1987, p. 113) describes ego development in terms of "a series of developmental stages that are assumed to form a hierarchical continuum and to occur in an invariant sequence...[that describes a] person's customary organizing frame of reference, which involves...an increasingly complex synthesis of impulse control, conscious preoccupations, cognitive complexity, and interpersonal style."

In this paper we describe limitations and some potentially serious problems with the ogive method and suggest an alternative method for aggregating the test items. One problem with the ogive or any cutoff method, or indeed any method that forces continuous phenomena into discrete values, is that they are sensitive to small variations and noise near the cutoff points. For example, a cutoff rule that assigns a TPR of "achiever" if 10 or more of the 36 items scores are at or above the achiever level is fragile for surveys that have 9, 10, or 11 such scores. In this range, if the score had just one or two more or fewer items at the achiever level due to rater error or inter-rater variation or an arbitrary glitch in the subject's thinking when taking the test, the score would jump an entire developmental level.

These issues are unlikely to impact the vast majority of the hundreds of published studies that use the SCT method, especially those with medium to large sample sizes. Small "random" variations that cause the TPR to "flip" up or down a level will "wash out" in statistical aggregations. Such issues might be more important for case studies and other small-N studies, where a single level difference in a couple of TPR scores could change the conclusions. Beginning with Cook-Greuter and continuing with Torbert and O'Fallon, the SCT is increasingly used to provide developmental scoring and coaching/consulting advice to individuals (something Loevinger strictly avoided). It must be said that all three of these researcher/practitioners and their associates are careful not to allow the use of the SCT for high-stakes assessment decisions such as hiring and firing (to the extent possible) but instead use it as a tool for self-reflection and personal growth (for individuals but also for teams and organizations). Nonetheless, the sensitivity of the cutoff method to perturbations is concerning since the characteristics of, and thus the advice given for, each developmental level is distinct (i.e. adjacent levels such as Expert and Achiever differ significantly along many dimensions).

The research in this paper was executed in an attempt to validate and improve O'Fallon's STAGES variation of the ego development theory and assessment, yet some of our conclusions concern the entire SCT lineage (other results apply only for the STAGES model). STAGES represents the most recently developed model in this lineage and differs from others by proposing a small set of underlying "drivers" or parameters (i.e., underlying mechanisms) purported to explain most or all of the surface features of the spectrum of meaning-making development. The research goals addressed in this report include:

- To further investigate suspected problems with the ogive cutoff method for aggregating item scores and to propose alternatives;
- To apply IRT and Rasch modeling to determine psychometric aspects of the STAGES SCT that were previously unaddressed in SCT research, including:
  - An item-based determination of overall test strength (including factor analysis and reliability);
  - Item-based analysis to identify items that behave differently than others or have sensitivity to particular regions of the developmental spectrum;
  - Test-length analysis to determine the validity of shorter versions of the SCT;
  - Scale uniformity and resolution to determine whether the categories (stages) in the model are relatively uniform and well differentiated;

- To assess other statistical properties of the measurement, including the skew and kurtosis metrics, indicating whether the spread of items in the surveys is normally distributed (on average).

## Project Beginnings

There were two initial motivations for this project. First, O'Fallon's statisticians suspected some fundamental errors in the assumptions behind the ogive cutoff method for aggregating the (36) response scores used to produce the total survey score (TPR) traditionally used in Loevinger-lineage sentence completion tests. One goal was to clarify these issues and identify alternatives to the ogive method.

Second, the main alternative framework to STAGES and other SCTs include the LECTICA assessments (based on hierarchical complexity theory (HCT), see Commons et al., Dawson et al., and Fischer et al.), and research in this lineage uses Rasch modeling for evaluating psychometric validity.<sup>3</sup> Rasch analysis is considered by most to be a subset of IRT methods, and IRT methods represent a more robust and modern set of methods than the classical test methods Loevinger chose over 30 years ago and continue to be used with the SCT. One of our goals was to upgrade methods to more modern techniques and also to determine whether IRT can shed light on how to replace the ogive aggregation method.

Over the course of this research, Murray engaged the assistance of many professional statisticians/psychometricians. Each result shown here was read and approved by at least two of these experts, although consultant reviews regarded checking that the methodology fit the context and conclusions and none checked the actual data processing steps (i.e., the author is wholly responsible for any such errors, and none of these consultants can be said to have vetted and agreed with this report in its entirety and full detail). We would like to thank the following for their help at various stages of the project: Matt Johnson – Columbia University faculty in psychometrics and statistics research; Trevor Bond – faculty at James Cook University and lead author of *Applying the Rasch model: Fundamental measurement in the human sciences* (Bond & Fox, 2001; Bond, 1994); Bob Dolan – a psychometrician in the field of educational technology who has worked for several publishing and research companies; Mike Linacre – psychometrician and developer of the WINSTEPS software package for Rasch analysis; Larry Price – Texas State University Professor and author of *Psychometric Methods: Theory into Practice*; and Moni Blazej Neradilek and Nayak Polissar – lead statisticians at Mountain-Whisper-Light Statistics, the company O'Fallon uses for much of her work on the STAGES model.

Part of this work was funded by the Developmental Research Institute (DRI), a registered non-profit research organization founded originally by Terri O'Fallon, Suzanne Cook-Greuter, and others, which is currently lead by O'Fallon, John Kesler, Venita Ramirez, and Judy Stevens Long. Although the mission of DRI is to further research adult developmental assessment and theory in general, we acknowledge that most of its board members have close affiliations with

---

<sup>3</sup> Kegan's model is also predominant in the field, but its assessment method, the subject-object interview, is much less frequently used, because it is much more time consuming to both administer and score.

O'Fallon, as does the author, and that there may be some unintentional bias toward the STAGES model in the reported research.

## SCT History

Below we describe the evolution of the SCT and compare its various versions to illustrate their significant similarities for two reasons: to argue that (1) many conclusions made in this study with STAGES assessment data will apply to other SCTs; and (2) prior research using the WUSCT, which constitutes the vast majority of the roughly 400 studies on the SCT, applies to the other SCTs.

The SCT for adult development originally developed by Jane Loevinger was later extended by Cook-Greuter and associates, Torbert and associates, and O'Fallon and associates. Although Loevinger remained an academic, the others branched out of academia to start companies that engage in developmental assessments, organizational consulting, and consultant trainings (Torbert works both in academia and industry consulting). Cook-Greuter was a student of Loevinger and Torbert was one of her dissertation advisors, and both Torbert and O'Fallon were at various times working as business colleagues of Cook-Greuter, who is currently a principal at Vertical Development Academy; Torbert worked at Harthill Consulting for 20 years and then founded Global Leadership Associates with Elaine Herdman-Barker; and O'Fallon is co-founder at STAGES International. All four of these organizations continue to work with the SCT and, as far as we know, represent the only organizations that currently provide scoring of and certification in the SCT. Vertical Development Academy calls their SCT "MAP," Harthill Consulting calls theirs LDP, Global Leadership Associates calls theirs GLP, and STAGES International calls theirs STAGES (Torbert et al. 2009, 2014 describes validity research on MPA, GLP, LDP).

Loevinger's model and assessment was extended in three ways by those that followed her: (1) the underlying psychological theory evolved, (2) the SCT itself was altered regarding the number and selection of stems, and (3) the analysis method defined in a "scoring manual" also evolved.

In terms of the SCT instrument, MAP and the "general protocol" of STAGES use 36 stems, following Loevinger, while LDP and GLP contain as few as 24 stems. Although each SCT variation has altered some of Loevinger's original sentence stems, all share at least 75% of the stems with the original WUSCT and with each other.<sup>4</sup> Although Loevinger and other researchers using her methods have validated and used an 18-item ("split half") variation of the SCT (see Novy & Francis, 1992; Holt, 1980), the short version is not used for the individual scoring currently done by the others. However, O'Fallon has been developing and validating "specialty" protocols with both varied lengths and 6 or more stems from the "generic" SCT replaced by stems focused on specific themes (e.g., love, money, spirituality, etc.). The research reported here utilizes the STAGES generic protocol.

---

<sup>4</sup> See Torbert 2014 for a detailed comparison of the MAP, GLP, and LDP variations of the SCT. Torbert and Cook-Greuter modified the WUSCT stems to "omit a number of gender-based items and, includes work or leadership-related stems" (Torbert & Livne-Tarandach, 2009).

In terms of the theoretical model, Cook-Greuter extended Loevinger's system by differentiating two stages (which she called Construct Aware and Unitive) within Loevinger's final stage (a model also used by Torbert and associates). O'Fallon's research has taken this progression further by differentiating four stages (5.0, 5.5, 6.0, and 6.5) within the same late-stage territory. O'Fallon also differentiated the Conformist (or Diplomat) level into two levels based on theoretical principles along with her other modifications. (Thus, STAGES defines 12 levels, MAP/GLP/LDP defines 9, and WUSCT defines 8). All of the scoring methods have procedures for marking a developmental stage as early, middle, or late, so the potential number of "strata" or sub-stages has a finer resolution. Although the early/late distinction is used in qualitative consulting with the client, for the research reported here all score data regards the per-stage resolution (i.e., early/late designations are not used).

Cook-Greuter attempted to address her notion that Loevinger's model was "lacking an underlying structural logic" (1999, p. 76) by linking the sequence of stages to "person perspectives" (p. 77) (e.g., first-person perspective, second-person perspective, etc., sometimes referred to as worldviews). This strengthens the ego development construct from a "soft" description-based construct towards a "hard" model-based construct, which provided a coherent theory that could support the entire developmental trajectory using a single concept (pp. 72–76). However, this change was applied only to the conceptual description of the model rather than to the scoring methodology. O'Fallon extends this innovation by more fully integrating the concept of person-perspectives into both the theory and the scoring procedure.

A scorer using the WUSCT, MAP, GLP, or LDP tests consults a scoring manual comprising thousands of example sentence completions organized by stem, level, and theme. The scorer attempts to match a sentence completion with an example or thematic group of examples, and if no match can be found, more general heuristics defined for each level (independent of stem) are used. Westenberg et al. (2004a, p. 485) note that "the scoring manual for the [WUSCT] consists of over 2000 response categories...about 80 response categories for each of the [...] items." The 300+ page scoring manuals used for the MAP, GLP, and LDP contain 16,000+ examples, and training to score from the requires many months and close supervision until acceptable inter-rater reliability is achieved (vs. a master scorer). For the two highest levels added by Cook-Greuter, scoring relies more on heuristics than exemplars (the method is used for the MAP, GLP, and LDP).

While all the other methods follow Loevinger in using an exemplar-based scoring method, the STAGES model differs by basing its model and scoring method on an underlying quasi-causal model that applies a repeating pattern using a small set of foundational parameters. The scoring procedure (the manual) is thus stem-independent and does not rely on exemplars but on underlying language properties and structures.

## The STAGES Model

Although the example-matching method was chosen by Loevinger, it entails drawbacks. Because human language is diverse and expressive and there are so many ways that an individual at level X could respond to stem Y, it is quite time consuming to add new levels or sentence starters, which requires the collection and validation of excessive amounts of data to define

exemplars and heuristic rules (e.g., see Minard, 2009; Cook-Greuter (1999)). For scorers, matching to exemplars can be tedious and time consuming, and these issues motivated O'Fallon to develop an alternative theoretical and scoring framework. Following years of scoring experience using the MAP framework and using insights sourced in Ken Wilber's Integral Theory (the AQAL model, 1995), O'Fallon proposed that a small set of underlying parameters could be used to explain (and score) the entire developmental sequence. This idea, in its final form after several iterations, has proven to be a valid method to understand and assess ego development.

STAGES defines 12 developmental levels (stages) in three tiers (Concrete, Subtle, and Metaware) with each four levels each. The levels in each tier progress through a sequence of four learning-and-acting orientations: Receptive, Active, Reciprocal, and Interpenetrative. The sequence is also defined as progressing through 6 "person perspectives," each of which has an early (passively oriented) and late (actively oriented) phase. The STAGES are thus identified as 1.0, 1.5, 2.0, 2.5...6.5, where the "1." to "6." reflect the six person perspectives and the ".0" and ".5" reflect the early and late phases of each person's perspective. Appendix 2 contains figures illustrating the STAGES levels, and for a detailed description of the STAGES model, see other articles in this journal special issue and material at [stagesinternational.com](http://stagesinternational.com).

Because it does not rely on exemplars but rather on underlying language properties, it is much easier to create modifications to the STAGES SCT compared to the other variations. The same scoring heuristics are used regardless of the sentence stem, and new stems can be added to the SCT without the need to conduct extensive research into how people might respond to the prompt. O'Fallon has developed several "specialty protocols" using alternate sentence stems, and in each case the internal consistency of the SCT remains high. In addition, because it is based on language properties rather than specific exemplars, STAGES is less susceptible to drifts in word meanings over generations and should be easier to apply to multiple languages. Finally, using generic theory-based scoring rules allows the method to be extended beyond the SCT to preform developmental analysis on arbitrary texts such as reflective essays, story narratives, speeches, and books (O'Fallon has applied the method to such texts, but we lack validity studies on these novel applications). (Note that, despite the notes above on the flexibility of the STAGES for creating alternative SCTs, the studies reported in this paper regard the standard-length "general" protocol, which has significant overlap with the stems of the other SCT variations.)

STAGES is also a fully "hard" stage theory in that its stage boundaries are determined by theoretical rather than empirical or practical considerations. The number of stages and their boundary lines changed several times during Loevinger's work, and Cook-Greuter made additional modifications based on empirical or pragmatic reasoning. Loevinger and Cook-Greuter explicitly chose a data-driven rather than theory-driven foundation for the ego development model, meaning they wanted the model to be driven by real observations rather than by assumptions imposed by an abstract theory. It is however difficult to claim that one has the correct or best stage definitions under such circumstances.

Data-driven methods are best for new and emerging theories, but at some point theory-driven approaches become feasible when enough is known to propose a stable set of underlying mechanisms that give rise to surface features of the data (the neo-Piagetian developmental

models of Commons and Fisher are also theory driven). The number, sequence, and boundary conditions of the STAGES model directly originate from its theoretical assumptions and will not change any time soon. Of course, science evolves through an iterative dance between theory (ideas) and evidence, and even set-in-stone theories must be occasionally re-evaluated in light of new data. O'Fallon readily acknowledges that a day will come when STAGES is superseded by another model – personal communication.<sup>5</sup>

The above paragraphs describe how STAGES differs from other SCT variations, including its purported benefits. Although this allows coordinating the results of our study with other SCT variations, it is important to note that for the purposes of the study reported here, STAGES is largely similar to its predecessors.

The first proof of this is argued in detail in "The Validation of a Novel Method for Assessing Developmental Stages of Meaning-Making Across the Life Span" (O'Fallon et al., 2020), which reports on a "replication study" of its concurrent validity showing that, for stages up to 4.5/Strategist, scoring with the STAGES scoring methodology correlates exceptionally well with Cook-Greuter's MAP method (and presumably the other SCT variations which are more similar to MAP). For stages above 4.5/Strategist, the definitions of levels diverge too much and there is insufficient data available to perform a proper comparison. However, from Cook-Greuter (2013), we can estimate that approximately 2% of the population or 7% of professionals are above the 4.5 level, so the replication results apply to the majority of people.

Second, it can be argued that the SCT has strong reliability as a methodology regardless of the mentioned variations.<sup>6</sup> This was reported in "The STAGES Specialty Inventories: Robustness to Variations in Sentence Stems," where O'Fallon created about six specialty protocols with alternative sentence stems. These all show high internal consistency via the Cronbach's alpha measurement, indicating that the SCT method is robust over many possible sentence stems (assuming they are rationally designed). In addition, as illustrated later in this paper using a "Cronbach-Mesbah curve," the SCT continues to have strong consistency even for 10 stems (.92), and even at 5 stems the metric is good (.85). In addition, the IRT and Rasch analysis reported later shows that all the 36 stems but one are strongly correlated with each other.

This all indicates that the SCT is a strong and valid methodology regardless of its variations and that certain results from one variation should be applicable to others. This claim weakens for

---

<sup>5</sup> We should also note that, because the scoring rules (manuals) are different, the territory for each level defined by the STAGES model need not map *exactly* onto the same territory in the prior models. Our understanding of the psychological and cognitive characteristics of each developmental level contains many aspects (e.g. ability to utilize to feedback, use of either-or vs. both-and language, and how past and future are conceived), and persons identified at level X in STAGES may, on average, have slightly different characteristics vs. persons identified at that same stage in another model. What *has* been empirically demonstrated (O'Fallon et al., 2020) is that, with all of these characteristics taken together as a whole to describe each ego development level, the STAGES model tracks very well with the other models (up through 4.5/Strategist).

<sup>6</sup> Statistical reliability refers to the consistency of a measure, i.e. whether it produces similar results under consistent (accurate, reproducible, and consistent) from one testing occasion to another, having acceptably small random errors. Overall reliability can include inter-rater reliability, test-retest reliability, and internal consistency.

developmental levels at the upper and lower extremes where there is less empirical data and more tentative theoretical understanding, but we do not yet have a sense regarding how much it weakens.

## Prior SCT validations

The literature on the SCT is extensive and includes over 40 years of meta-analyses and critical overviews, substantially supporting its validity and usefulness (see Cohn & Westenberg, 2004; Manners & Durkin, 2001; Holt, 1980; Novy & Francis, 1992; Jespersen et al., 2013; Westenberg et al., 2006; Forman, 2010). The paper "Sentence completion assessments for ego development, meaning-making, and wisdom maturity, including STAGES" (Murray, 2017) contains a lengthy summary of such studies, illustrating the validity and reliability of the SCT and confirming Westenberg et al.'s conclusion that "findings of over 350 empirical studies generally support critical assumptions underlying the ego development construct" (p. 485).

Blumentritt (2011, p. 153) says that "more than 1,000 articles and book chapters have been published examining nearly every conceivable aspect of the construct and measurement of ego development," overall showing "substantial support" for the theory and measurement. Most of these studies are based on the WUSCT, but some use the MAP/GLP/LDP variations. For the reasons given above, we claim that these results apply to STAGES as well, and results of the meta-analyses include:

- "Psychometric studies of the WUSCT...invariably report high levels of inter-rater reliability" (Westenberg et al., 2004a, p. 603; and see Torbert & Livne-Tarandach, 2009 for strong results on the MAP, GLP, and LDP; and O'Fallon et al., 2020 for strong IRR on STAGES).
- "The WUSCT [displays] high internal consistency: Most studies report a Cronbach's alpha of .90 or higher" (Westenberg et al., 2004b, p. 693; and see Torbert & Livne-Tarandach, 2009 for strong results on the MAP, GLP, and LDP; and later in this paper for strong internal consistency in STAGES).
- "In terms of test-retest reliability, when sufficient time is allowed between the two tests to allow for motivational effects, significant correlations have been found between test and retest scores" (Manners & Durkin, 2011, p. 545).
- "The face validity of the SCT is demonstrated by the sheer fact that it has been used in more than 300 research studies [including] such diverse topics as parenting behaviors, managerial effectiveness, and the effects of meditation on recidivism rates" (Phaffenberger, 2011, p. 10).
- The SCT has "excellent reliability, construct validity, and clinical utility" (Westenberg et al., 2004a, p. 596).
- Longitudinal studies have confirmed the sequential invariance of the developmental steps (i.e., no stage can be skipped) (Loevinger, 1998).

- The SCT has incremental validity over IQ and SES measurements (e.g., Browning, 1987; Cohn & Westenberg, 2004);
- The SCT has proved applicable in different countries, cultures, and languages (see Carlson & Westenberg, 1998).
- Although there have been fewer studies on the predictive validity of the SCT, Vincent notes that "a growing body of studies is showing associations between increasing consciousness development and better leadership performance and organizational outcomes" and cites a substantial 21 articles in this regard (2015, p. 2). Other indications of predictive and external validity can be found in (Torbert, 2014; McCauly et al., 2006; and Harris, 2005).

Torbert (2014) summarizes a number of studies on the MAP, GLP, and LDP instruments that arose after the WUSCT.

Our study adds to these findings about the SCT (confirming a few of them) and focuses on two areas that have been under-researched: item-level statistics and methods for aggregating the SCT items into the overall score.

## The Ogive and TWS Aggregation Methods

How to calculate the overall score of the SCT turns out to be a complex question. One reason is that participants are expected to exhibit a range of developmental levels in their answers. In a projective test the subject is responding freely, rather than trying to score as high as she can, and it is actually viewed as more healthy or well-rounded when responses range over at least 3 or 4 levels (however for individuals at low developmental levels there is less room to range over). A survey with all items within a tight range of 2-3 levels is not only unusual, but it is considered possible evidence for a rigid personality style.

Most psychometric tests take the simple sum or mean over their items, although sometimes the items are weighted according to difficulty or information power. This is however not a particularly satisfying method for the SCT. For a projective test where the subject is not attempting to "answer" or "solve" anything, some items may demonstrate the subject's true competence while others may naturally be more "off the cuff," meaning short and simple, playfully carefree or irreverent, or even impulsive or unreflective "outbursts." Our intuitive understanding of the construct and the test seems to demand that we focus on the highest scores that the subject produces.

For example, if someone attains a score of, say, 4.5/Strategist level on 8 of the 36 items, then they must have a strategist level of meaning-making, because according to both the theory and empirical evidence, it is difficult to "fake" or "guess" items to produce a score higher than one's actual developmental level (Redmore, 1976). Whether the other scores on the test are near 4.0/Pluralist or 2.5/Conformist does not change the fact that they seem to have achieved a 4.5 developmental level. However, having the rest of the scores at 4.0 vs. 2.5 *does* strongly affect the

average or sum of the items. One of the justifications for using the SCT cutoff method is that it prioritizes higher scores (or more accurately scores that are rarer within the population).

Within the Loevinger tradition there are two established methods for aggregating the 36 SCT items to obtain an overall score for the survey: the ogive and the TWS. The most commonly used by far is a cutoff method called the ogive method (yielding the total protocol rating, or TPR, sometimes called "center of gravity"). In the ogive method, one follows a procedure to check each level in turn to determine whether there are sufficient scores at or above that level. The ogive cutoff method essentially "throws out" data beyond the cutoff. According to Cook-Greuter (1999, p. 222), "One thus rates a test at the highest stage that is exhibited if enough evidence at that stage is present and not enough at the next higher level." The cutoff number varies over the levels, as detailed in Appendix 1.

Holt et al. describe Loevinger's reasoning: "Human development has this odd, psychometrically inconvenient property of maintaining the potentiality to respond on many lower levels after one has, in a certain sense, left them behind. The ogive rules respect this peculiarity and allow for it" (Holt et al., 1980, p. 917). 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 much less frequently used regards the total weighted score (TWS) method,<sup>7</sup> which sums the scores over the 36 items and returns an integer typically in the 300-600 range. The formula used for STAGES is:  $(f_{1.0} * 2) + (f_{1.5} * 3) + (f_{2.0} * 4) + \dots + (f_{6.5} * 13)$ , where for example  $f_{1.5}$  is the count (frequency) of completions scored at 1.5.<sup>8</sup> There are slight variations on the constants used in this formula for different frameworks.

The ogive method is preferred because Loevinger believed that its categorical result is more meaningful and has more construct validity compared to 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.<sup>9</sup> Researchers in the psychological and social sciences almost universally acknowledge that psychological phenomena are complex and that using discrete

---

<sup>7</sup> 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 per-item 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.

<sup>8</sup> 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.

<sup>9</sup> 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.

categories involves substantial simplification which is thus performed for practical reasons. There is considerable debate within the developmental psychology community regarding whether development occurs 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 represents a practical and/or theoretical decision but is unsupported from an empirical perspective.

Although the discrete ogive measurement is predominantly used, the continuous TWS measurement is sometimes used for comparative, pre-post, or longitudinal studies since it is more sensitive to small changes. It can require 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 (Lee, 2012) and concern how much information is needed to conclude that a person is at level X with for example 95% confidence.<sup>10</sup> (95% is usually chosen as the confidence factor to use.)

As noted in Appendix 1, the cutoff procedure begins with the top level and proceeds to check each lower level until 3.0/Expert, at which point it switches to testing the lowest level and works its way up to 2.5/Conformist (the same method is used for Loevinger and Cook-Greuter with slightly different stage sequences compared to STAGES). This follows the principle of first checking for the least likely possibilities. The 2.5/Conformist level was determined to be the "average" or most frequent level (the "prior probability" in Bayesian terms), 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 guessed given no other information.

Note that in actual scoring for clients according to both the STAGES method and the earlier WUSCT and MAP methods, the analyst can adjust the final total score based on global subjectively determined factors. An ogive score, especially one on the edge of a cutoff value, can thus be tweaked up or down a level given sufficient evidence from considering the big picture over all stems (remember that each stem is scored by itself without influence of other stems or the whole). Our analysis estimates that this TPR adjustment is performed less than 10% of the time.

## Problems with Ogive and TWS Aggregation Methods

Here we arrive at one of the main purposes 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. Given the large number of studies using the SCT, the ogive method for accumulating items is surprisingly rarely if ever questioned or analyzed. One reason for this may concern the large number of studies in the field, as 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.

---

<sup>10</sup> Based on the idea that, as in medical diagnosis, the rarer a symptom, the greater its predictive power.

We will mention problems with cutoff methods and problems specific to the ogive cutoff method, specifically:

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 problem with the cutoff method is that, based on Bayesian probability concepts, it makes assumptions about the overall population, which entails three problems:

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 Conformist as being the average (i.e., the prior probability) was an educated guess based on a number of studies (each with subjects in the hundreds at most), however many of these studies regarded populations such as pregnant mothers, prison populations, and others with below-professional educational or socioeconomic status. This may be a reasonable guess for the average U.S. population, but the modern uses for the SCT predominantly concern professional or self-improvement contexts where "leadership skills" and "self-reflection skills" are considered important. Articles by Cook-Greuter, Torbert, and O'Fallon have tried to estimate the percentages of each developmental level within specific populations such as the general population, college-educated, professionals (managers), and consultants (see Cook-Greuter, 2002; Torbert & Livne-Tarandach, 2009), but none of these are based on random samplings of populations.
2. Within any specific population, the base estimates ("prior probabilities") will differ, and thus the Bayesian assumptions differ and therefore the cutoffs should ideally be modified, but this is never performed. In practice, the lower and middle cutoff numbers, including the procedural "switch-over" after 3.0/Expert from checking downward to checking upward, are never modified. This is probably because one desires results to use standard methods and to be comparable with past research in the field. In addition, gathering

enough data in any sub-population to determine the proper new cutoffs is usually prohibitively difficult.

3. Several sources indicate 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 when retaining 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 or 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" (p. 121). Cook-Greuter (1999) 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. (p. 228)

Cook-Greuter chose stricter cutoff criteria for the top levels compared to Loevinger and 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 in or near what Loevinger considered the "extreme" of the upper end, and thus these problems are central to modern assessments rather than being relinquished to rare occurrences. Cook-Greuter and associates have recently adjusted their cutoff values and no longer rely on Bayesian reasoning (personal communication) but continue to use a similar cutoff scheme, which is thus expected to include most of the issues mentioned here. When adding three new levels to the scheme for the STAGES model, O'Fallon had to take an educated guess for the cutoff numbers of these new levels (each of which split a prior level in half).

Most statistical metrics involve some quasi-arbitrary parameter or parameters (e.g., when using p-values to estimate statistical significance, the significance cutoff alpha is chosen as 5% or 1%, etc.). However, the degree of arbitrariness in the ogive method is considerable given the methods for determining the cutoffs and the fact that there are at least as many such arbitrary parameters in the method as there are levels.

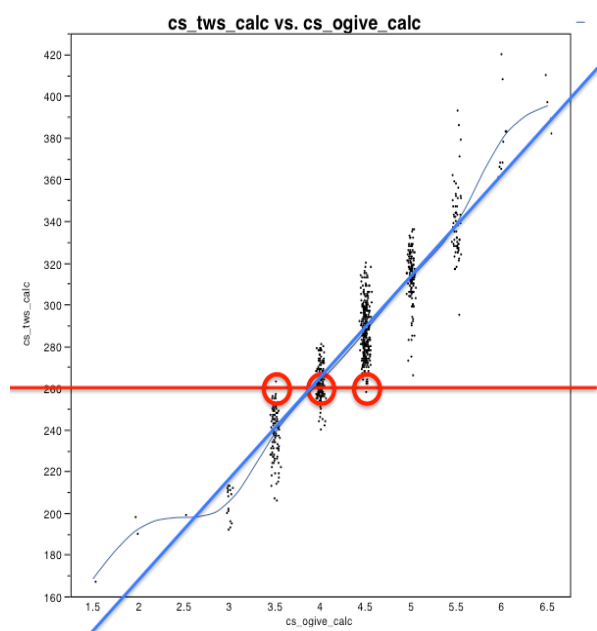
**Lower cutoffs confuse development with "shadow" evidence.** The scoring rules, including the examples given in the scoring manuals used for WUSCT and MAP, seem to confuse the immature development of ego with "shadow-crashes" or momentary regressions to lower ego levels that may be triggered based on the subject of the stem. Because the levels 1.0/Impulsive and 1.5/Opportunist are "rare" in normal adult populations, they are checked before 2.0/Rule-oriented and 2.5/Conformist 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 have their center of gravity (TPR assigned to those levels

respectively. Achieving such a low rating can result from a simplistic one-word answer but it can also result from a seemingly narcissistic or impulsive answer, including the use of vulgarity or mentioning of violence. We argue that the scoring method confuses actual level of development with "shadow crash" phenomena for lower-level responses.

For example, a person may have 19 stems at 3.5/Achiever yet will achieve an ogive score of 2.0/Rule-based if they have only 7 stems at 2.0 or lower., and such a person's center of gravity should clearly be closer to 3.5 than 2.0. Because there are very few people with a center of gravity below 2.5 in the populations that currently use the SCT, this problem has not been salient, however one area where it is limiting regards studying children. O'Fallon wants to use the STAGES model (and has already collected data) for pre-adult populations, and in such cases one desires 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 represent an additional reason for transitioning from the cutoff system. This suggests having separate measurements for developmental level (e.g., a sum or average score like the TWS) vs. a shadow-related measurement (e.g., counting the number of low-level, more impulsive, or reactive scores). We later propose such a system.

**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. For example, if the cutoff for some level is 10 items and a survey has 9, 10, or 11 items at that level, then a small difference in item scores such as due to scorer error or to a person not expressing themselves "at their best" in the moment of completing the stem would



**Figure 1.** Ogive vs TWS.

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 to ameliorate this issue by specifying early, mid, or late within the stage, but most trained scorers do not perform 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, which plots TWS vs ogive values for each survey in the STAGES database (described in the Methods section; the TWS vs ogive graphs for all of the SCT variations will have similar properties). For example, 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 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 such as 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, intuitive, and standard summation statistic used in most psychometric research, and this peculiar behavior is a direct result of using arbitrary cutoffs at each level.

**Additional issues with the TWS multipliers.** The progression of weights used for the standard SCT TWS calculation is somewhat arbitrary as well. 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), which makes the TWS very nearly like a summation but not exactly the summation of the items.<sup>11</sup> The integer progression of multipliers assumes an equal separation between each level. However, because Loevinger and Cook-Greuter occasionally modified the stage number and boundaries of their models based on data-driven and pragmatic factors, essentially changing the "joints" in how the spectrum was "cut," this assumption cannot be made.

Nothing in the theory or data analysis could address the issue 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, which made revised TWS values incompatible with prior results. We replace the ogive cutoff with a system that uses sums/averages in a way that does not have arbitrary multipliers and will not have to be modified as the model/theory evolves. We also show the analysis of equal spacing in the STAGES levels.

**Errors in the ogive formula and assumptions.** 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 and the Bayesian reasoning used. It seems that of the hundreds of studies building on Loevinger's work, none 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, however when desiring a strong foundation for the SCT, these errors represent another reason to transition beyond the ogive cutoff method. The most affected conclusions of past studies are probably those that make inferences to the general population. These types of inferences represent a very small percentage of the results of published studies, which are usually seeking correlations between ego development and other phenomena (conclusions which remain largely valid in light of the errors we will mention) rather than making claims about whole populations.

First, the formulas used by Loevinger seem to assume independence among the stems, which is not the case. Second, although the ogive method is inspired by Bayes' theorem, it also seems

---

<sup>11</sup> For a simple summation, the multiplier of each level would be the order of that level, i.e. the frequency (number of) the first level would be multiplied by 1; the frequency of the second by 2, etc.; again, under the (problematic) assumption of equal level spacing.

to violate certain principles in Bayes' theory. In the ogive procedure (Appendix 1), if a cutoff values is satisfied (e.g., 4 items are at or above 4.5/Strategist) then all other information is disregarded (i.e., the other 32 stems). Central to Bayesian theory however is the idea that each new piece of information accumulates to inform the estimate of the resultant probability, meaning information is not discarded. Third, one of our consultants noted 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."

Finally, the equation used to calculate probabilities for any level is (from Cook-Greuter, 1999; and Loevinger & Wessler, 1970):

$$P_s = \frac{Pr * T_p^x}{Pr * T_p^x + (1 - Pr) * F_p^x}$$

Where:  $P_s$  = probability that a person is "at" some stage  $s$ ;  $x$  = number of positive signs for stage  $s$ ;  $Pr$  = the assumed proportion of the population at stage  $s$  (prevalence or prior probability);  $T_p$  = the assumed true positive rate of a single sign to correctly classify at stage  $s$  (sensitivity); and  $F_p$  = the assumed false positive rate of a single sign to incorrectly classify not at stage  $s$  (1 - specificity).

Consultants Polister and Blazej-Neradiek produced a report illustrating in detail that the ogive formulation: (1) tries to assume that the survey items are independent, which they are not; (2) makes an error in the probability algebra, which has significant impact on outcomes; and (3) that the values for parameters  $Pr$ ,  $T_p$ , and  $F_p$  are too arbitrary, subjective, and sensitive (i.e. small variations have substantial impacts).

The details of this report exceed our scope here (copies of that report are available from Murray). We present all of the above information as second-hand to report what our consultants found, as the author lacks the statistical skills to verify or reproduce these arguments from scratch. Such arguments are however not necessarily needed given that Cook-Greuter (in her 1999 dissertation) made significant efforts to critique Loevinger's strict use of Bayes' theorem to determine the cutoffs, especially for the higher levels. Although there was much hand-wringing in justifying the new set of chosen cutoffs, in the end they were arbitrary, as she says they are "arbitrary not in the sense of random or capricious, but in the sense of arbitrated, reasoned, deliberated." She provides a logical rationale for her cutoffs, but not a *compelling* or compulsory one, meaning that – i.e. another researcher might produce completely different cutoffs, using a valid but different rationale. The cutoff method was so standard an aspect of the SCT method that finding an alternative was, apparently and understandably, too radical a break with the tradition, and. So the "battles lines" of the methodology were thus drawn at *what* the cutoffs should be rather than *whether* a cutoff method should be used.

**Summary.** That there seemed to be significant 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).<sup>12</sup> In addition, another one of our consultants who took a cursory look at the ogive method, called it outright "indefensible." It is also true that no consultant could recall anything like this Bayesian-inspired cutoff method being used in test psychometrics in psychological or social science research.

## Research Methods and Results

### Method and Data Source

We include the method and results in the same section because we apply various statistical tools and software packages to a pre-existing dataset to answer our research questions. Some analysis was conducted using the R statistical package's popular IRT module (the lmt R library), while most used the Winsteps software (winsteps.com), which specializes in Rasch analysis (i.e., one parameter IRT) (Rasch, 1980; Andrich, 1988; Bond & Fox, 2001). The IRT and Rasch analysis are described in Appendix 3.

Our dataset originates from the software package and website that has been used for (1) subjects to take the SCT (survey) online and have their responses and demographic information stored and (2) for certified scorers to read a subject's responses to SCT stems and rate stems and the entire survey based on procedures outlines in the STAGES scoring manual. The data were collected and are owned by STAGES International (which was originally run by O'Fallon's prior organization Pacific Integral). We used data from the "general protocol" containing the standard 36 stems for the STAGES SCT. The data were anonymized to contain no identifying information and were stripped of the sentence completion text, with the result that the only relevant data for each survey item are the item score (1.0 to 6.5) and survey-level numerical data such as the TPR.

The data consist of 740 surveys taken by 644 clients (some took the test more than once) between March 2009 to July 2018 (i.e., over about 9.5 years). 330 clients were female, 272 male, and 42 of unknown gender. Education levels varied widely, but the vast majority were professionals with college degrees. The age frequency per survey is: 16-29: 49; 30-39:149; 40-49:186; 50-59:171; and 60-77: 115 (i.e., a relatively mature subject pool). Subjects were from a variety of locations around the world (mostly from the U.S, and EU but included Ethiopia, Australia, New Zealand, Canada, Russia, Kosovo, Pakistan, and China). All the participants spoke English as their first or second language.

### Descriptive and Basic Item Statistics

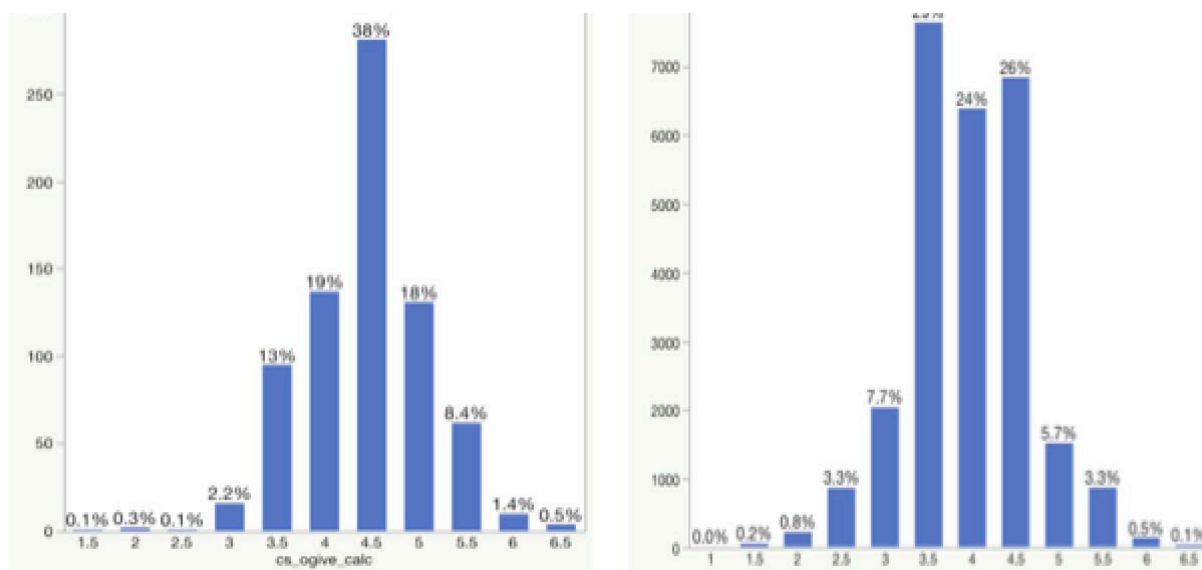
**Histograms.** Figure 2 shows histograms for the score at the item and survey (protocol) levels. (Note that this does not represent all of the data O'Fallon has from scoring, especially for the top two levels, as there are some small research projects that are not stored in this database.) Stage 4.5 surveys predominate, which indicates the types of clients and research subjects that come to STAGES International. Frequencies at the item level are high at 3.5, 4.0, and 4.5, which differs

---

<sup>12</sup> Note that we did not set out to, or ask these consultants to prove that the ogive method was flawed. At most we summarized arguments made by other consultants, and asked them to verify or argue with those.

from the survey level in large part due to how the ogive method focuses on the higher scores.<sup>13</sup> The IRT methods used in our analysis are robust with regard to such skews in data representation, so our conclusions should be valid for all levels except 1.0, 1.5, 2.0, and 6.5, which have very low representation.

To test whether standard statistic methods would apply to these data, we first investigated patterns of the 36 items within the survey and specifically the standard deviation of item scores and their normality (skew and kurtosis).



**Figure 2.** Histograms – Left: Survey ogive (TPR); Right: Item Scores for each level.

**Table 1.** Data shown in Figure 2.

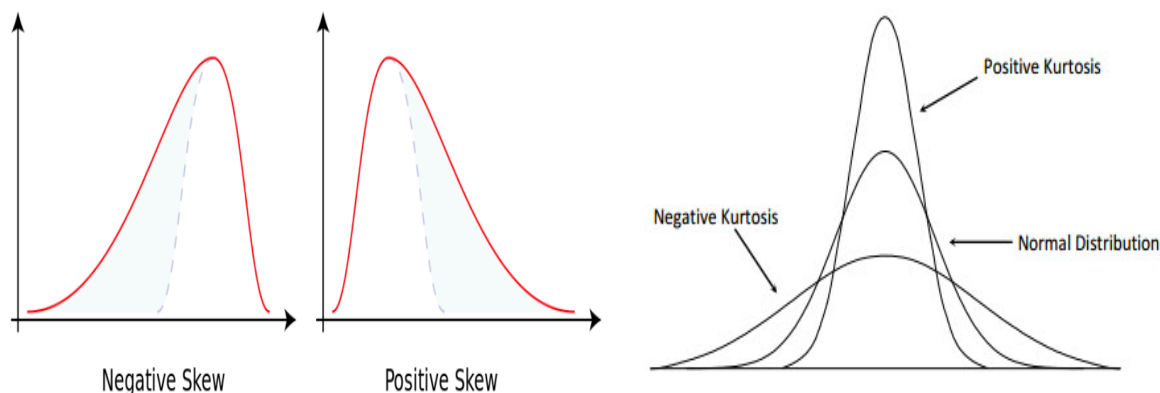
Stage:	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	Total
#surveys	0	1	2	1	16	95	137	281	131	62	10	4	740
#items	4	57	223	870	2040	7641	6388	6839	1522	874	146	36	26640

**Skew and kurtosis.** We checked the skew and kurtosis measures for items within a survey, which assess how close a dataset is to a normal (bell-shaped) distribution within each survey. This answers the question of whether items on average tend to "pile up" toward the earlier or later end of the scale. Figure 3 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 740 surveys, the mean kurtosis was 0.37 and the mean skew was 0.18, meaning on average the scores within a survey are normally distributed around the mean. This is a bit

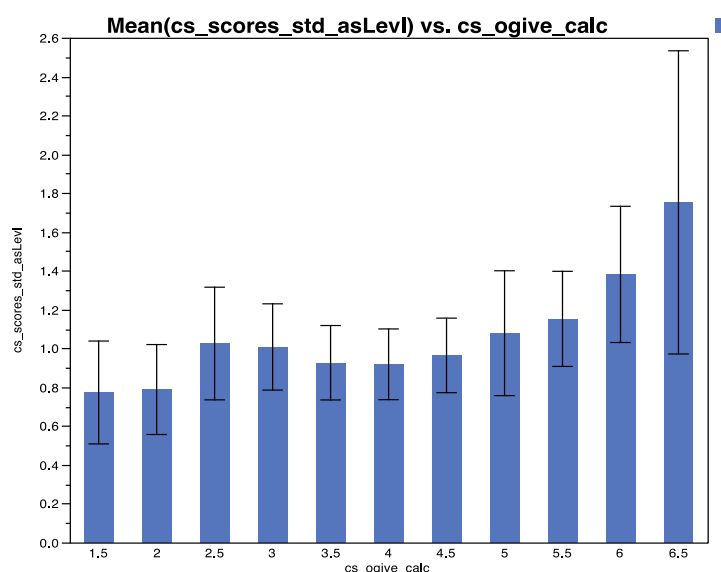
<sup>13</sup> Note the curious dip at 4.0 for item scores. We are still looking into why this might occur.

surprising since 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, which gave us confidence that more standard aggregation methods would be applicable.



**Figure 3.** Illustration of skew and kurtosis deviations from a normal distribution.

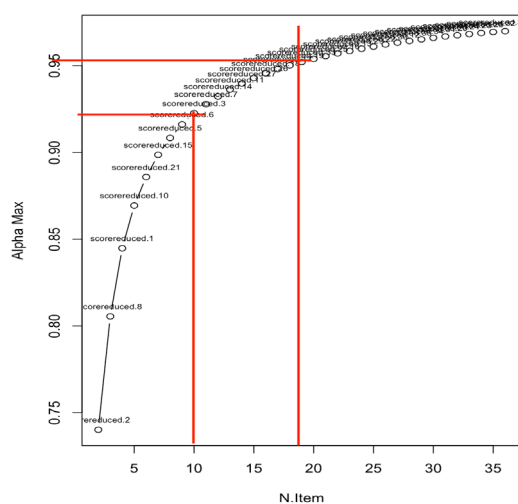
**Standard deviation.** We also examined how the standard deviation of scores within a survey varied based on the overall survey score (ogive). The results in Figure 3 show that the standard deviations for all levels (other than the top and bottom 1.0 and 6.5) are remarkably similar (.8 to 1.2). (The error bars in the figure show the variance in the standard deviations). The overall shape of the distribution of the 36 scores within a survey thus changes little for individuals at different centers of gravity, meaning that the spread of scores or width of the item's "bell curve" within a survey is surprisingly uniform regardless of the stage. There is however a slight upward trend, which would be expected because the higher the score, the more room one has to "roam" within a projective test.



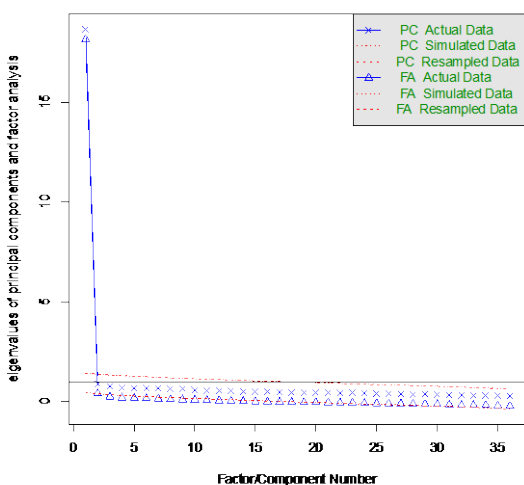
**Figure 4.** Standard deviation survey scores vs. stage.

These metrics also provide information about what the cutoff method is "throwing away" since it focuses on the highest scores (the number that depends on the level – see Appendix 1) and ignores the rest. As explained later, eventually we decided to calculate the total score using a standard number (6) of the highest stems. The normal distribution of the data and the fact that the SD shows the items cluster around  $\pm 1$  or 2 levels from their mean support the reasonableness of this type of aggregation method.

**Test length.** Figure 5 shows a Cronbach-Mesbah curve illustrating the Cronbach's alpha score for successively fewer test stems (with the worst correlated items removed first). The curve shows that for a half-item test, the reliability is still excellent (.95) and that even with 10 stems, the SCT has very good reliability (.92) and that even at 5 stems the reliability is quite good (.85). This supports any future projects that use surveys with fewer than 36 items (which are now being deployed by Stages International). From a purely statistical perspective, one might wonder why the SCT uses so many items (assuming this pattern was also true of the WUSCT or MAP variations of the SCT, although we lack item-level data for these). According to O'Fallon, the full set of 36 responses is important for identifying themes and patterns in the detailed reports and coaching sessions given for individualized SCT assessments (personal communication).



**Figure 5.** Cronbach-Mesbah curve.



**Figure 6.** Factor/Component Analysis.

the abstraction or complexity of the stem concepts. For example, "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. The figure also shows

**Factor analysis.** Figure 6 shows the output of a factor/component analysis of the data which, along with the numerical output, confirm the prior finding (for the WUSCT and now for STAGES) that the SCT loads on one factor and seems to be measuring one latent variable. This means that among the 36 stems, none seem to cluster together as if they were measuring a sub-skill of the overall capacity.

**Item difficulties and item standard deviations.** As a projective test, the SCT theoretically assumes that all items are approximately of equal "difficulty" on average. The stems are intended to "triangulate" or parallax the measurement from many life contexts (family, self, work, public life, etc.). Individuals are expected to vary regarding which life contexts evoke lower level responses (as in "shadow crashes"), but the item differences are assumed to be entirely personal rather than based on item difficulty. We can use IRT to test this assumption.

How different are the 36 stems? Figure 7 shows "High and Low Averages" and standard deviations per stem, and the tables are truncated to show only the highest and lowest thirds. The average score for each stem ranges from 3.71 to 4.21, which is a rather modest range indicating a small variability in stem "difficulty." Some of the stem difficulties correspond to what might be expected based on

that the stem "I am" has the highest standard deviation, which is consistent with the later IRT analysis showing it stands out as correlating the most poorly with the set of stems.

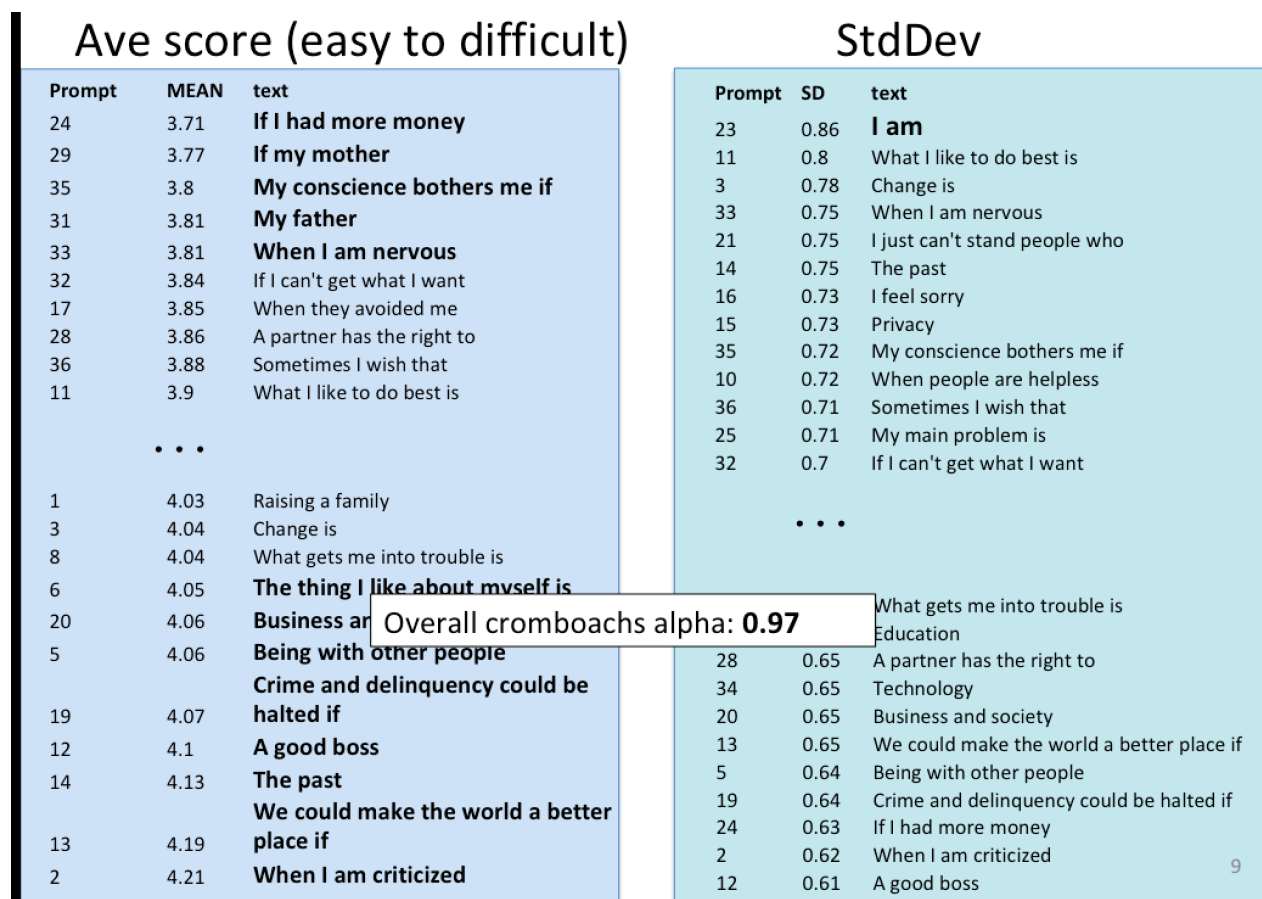


Figure 7. High and Low Averages and Standard Deviations for Stems.

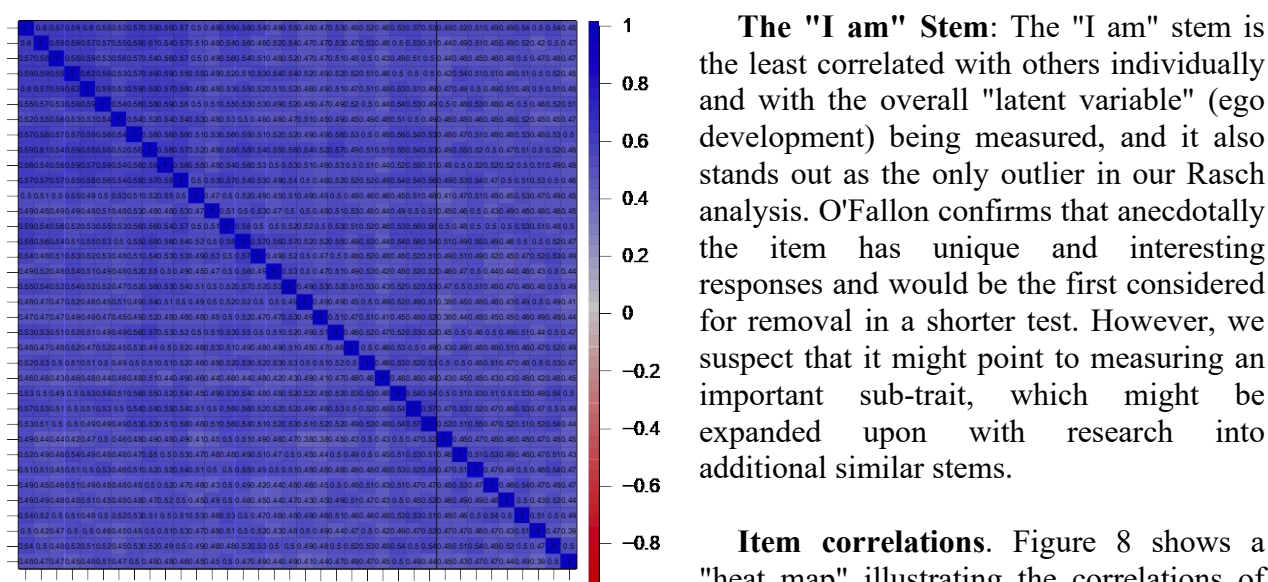


Figure 8. Item correlations Heat Map.

between 0.32 and 0.53. The Rasch analysis described later found all items except "I am" to "fit the model" of a rigorous test instrument.

## IRT and Rasch analysis results

We began our analysis using the R software's lmt package and then switched to the Winsteps program for more detailed Rasch analysis.<sup>14</sup> The two systems provide relatively equivalent answers for the tools they have in common, but each also has tools the other lacks. (Information about IRT and Rasch analysis is found in Appendix 3.)

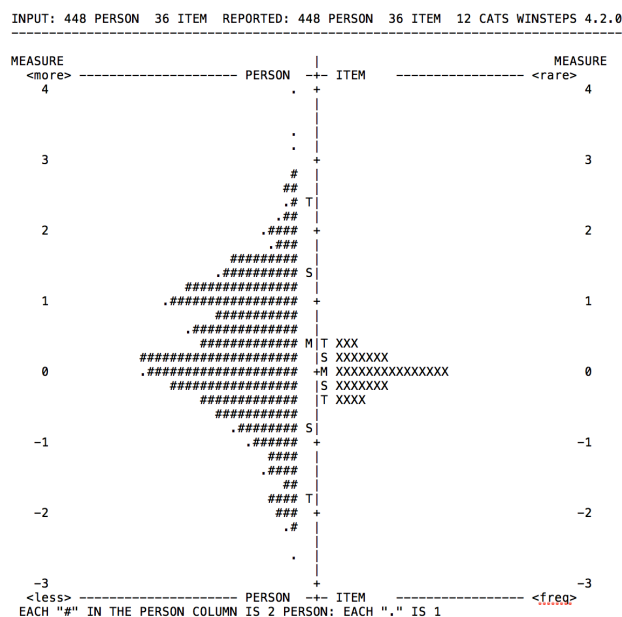
We tested the 1-, 2-, and 3-parameter versions of IRT modeling and found that the 1-parameter version (which the Rasch model uses) was sufficient. (The 2-parameter version had a slightly higher fit, but the fit of the measurement theory with our developmental theory led us to use the Rasch modeling.) Within the IRT analysis, we tried the generalized partial credit (GPC) model and the graded response model (GRM), finding that the simpler GPC model was sufficient (both the AIC and BIC variations of the GPC model were tried with little difference found). We also conducted some preliminary experiments using models outside of the IRT framework, including "mixture models," "grade-of-membership models," and structural equation modeling (factor-analysis-based item response trees) and did not find evidence that these more complex methods added value to our analysis. Within the Rasch/Winsteps analysis for polytomous items, we used Andrich's rating-scale model" (RSM) (the simpler of two alternatives since the data fit the Rasch model well). Information on how to interpret Rasch results including WINSTEPS outputs can be found in Linacre (2018), Bond and Fox (2001), and the [winsteps.com](http://winsteps.com) and [rasch.org](http://rasch.org) websites.

**Overall test strength.** Figure 9 shows the person and item distributions, representing the primary or most basic type of output from the WINSTEP (Rasch) analysis. The bar chart on the left shows the distribution of subjects (surveys) while 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 (technically from negative to positive infinity). 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 favorable. (One can see the dent in the curve representing lower occurrence at the 4.0 level mentioned in Figure 2.)
- The items are all clustered around a **similar difficulty** (the mean difficulty), as noted above. This is unusual for most tests, which are expected to have a range of easy and difficult items, but the tight range of item difficulties was expected for our projective test (see Figure 5).

---

<sup>14</sup> For the WINSTEPS analysis we limited the data to only those 448 surveys 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). Running IRT analysis with all 740 produced similar results.

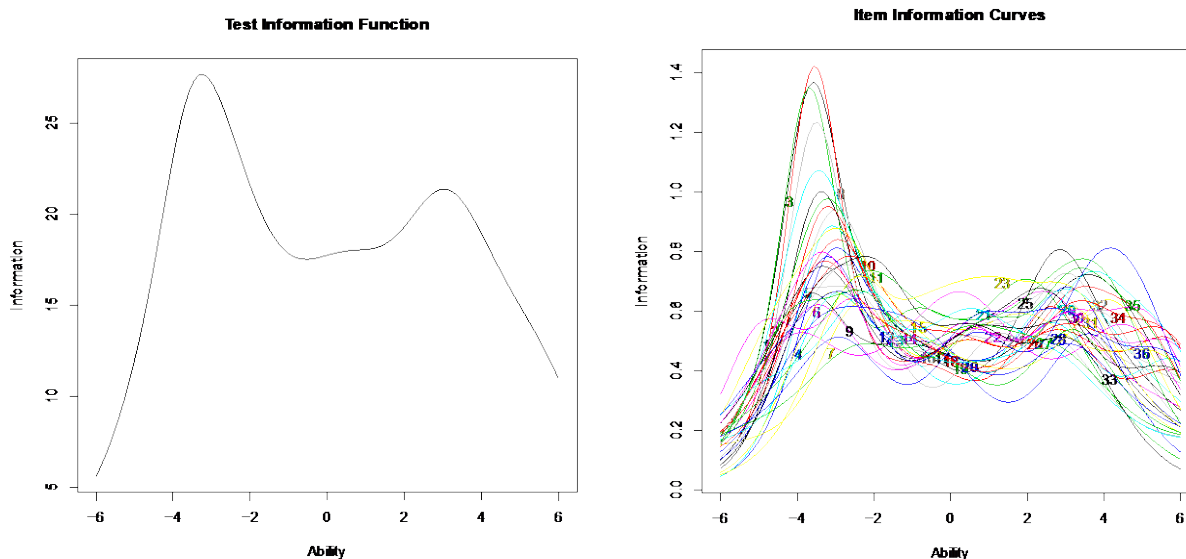


**Figure 9.** Person and Item Distributions from WINSTEP analysis.

Other output tables from the WINSTEP analysis confirm that the test would be just as powerful with fewer items. Many items were redundant from a psychometric perspective, but for the STAGES assessment in its normal use, the full set of 36 items gives scorers, debriefers, and coaches required additional information for the client.

**Item discrimination.** Figure 10 shows the individual item information curves and their combination in the test information curve," which illustrates that all items on the general protocol have *good overall discrimination*. Although there is a dip for the middle values, all of the values are acceptably high. Scoring is very *consistent* (i.e., good quality) across the prompts (similar to the Cronbach's alpha results). (It would be possible to conduct a deeper analysis on how each item performs

across the range of all levels, but this is not needed to answer our research questions.)



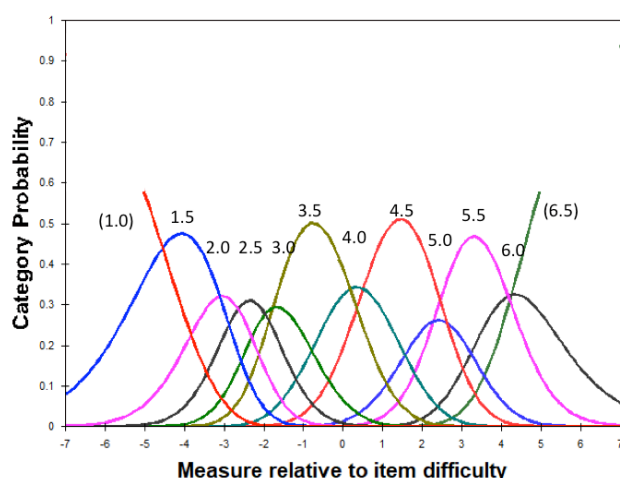
**Figure 10.** Test Information Curve and item information curves.

Figure 11 shows the probability category curve (PCC) which plots the probability of observing each ordered category according to the Rasch model relative to the overall item difficulty. For example, the "4.5" curve shows the probabilities of a person with an overall (average) 4.5 ability, which shows – for any given test item, what they are likely to score in terms of the latent variable (which is calibrated to 0 at the average score around level 3.8). They are most likely to exhibit about 1.5 (the peak, which is 1.5 standard deviations higher than the

average), and the probability of scoring higher or lower on an item declines on both sides as shown in the figure. The first and last curves (1.0 and 6.5) are always ignored in such plots (due to the math they always extend to infinity). The intersections of the curves for adjacent categories show the "Rasch-Andrich thresholds" for demarcating the range of each category.

The curve illustrates a generally robust test, however what is notable about the graph is that every other curve is (usually) lower. The lower curves show levels (response categories) that are not as clearly differentiated from their adjacent categories, and it appears that the X.0 levels are not as crisply measured as the X.5 categories. The scoring is strong at discriminating the 6-person perspectives but not as strong at discriminating early (passive) from late (active) halves of each person perspective. This makes sense based on the hierarchical nature of the model parameters, in which the tier distinction is the most crude, the individual/collective dimension is the next most refined, and the active/passive dimension is further refined. It is also possible that the observed pattern arises because the choice of items is better at eliciting active vs. passive responses.

O'Fallon hypothesizes that these dips for the X.0 levels are also because they regard "passive" phases in which the subject is beginning to recognize objects of awareness but is still struggling to find specific language for them, so they may drop to the prior level in terms of vocabulary.



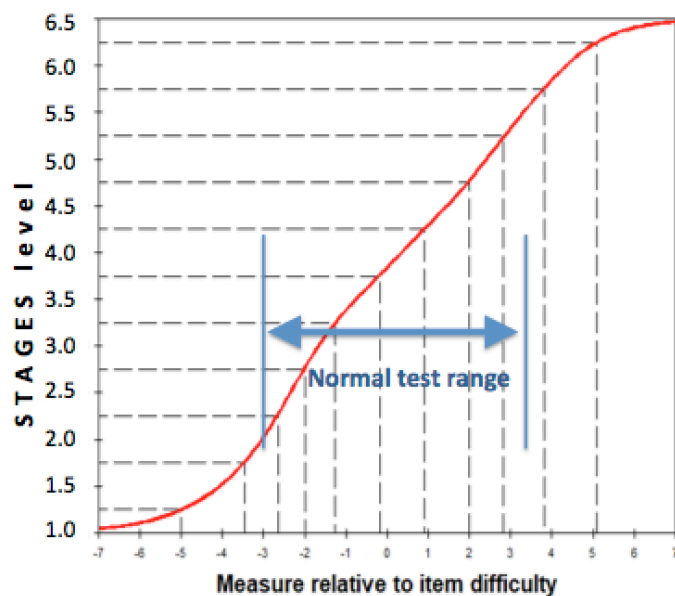
**Figure 11.** Probability Category Curves.

range of scores (2.0 to 6.0 at between +4 and -4 standard deviations from the mean, which is where 99.9% of item responses fall), and the spacing between levels is relatively uniform. The relative uniformity of spacing can also be observed in the cumulative probability curves in Figure 13.

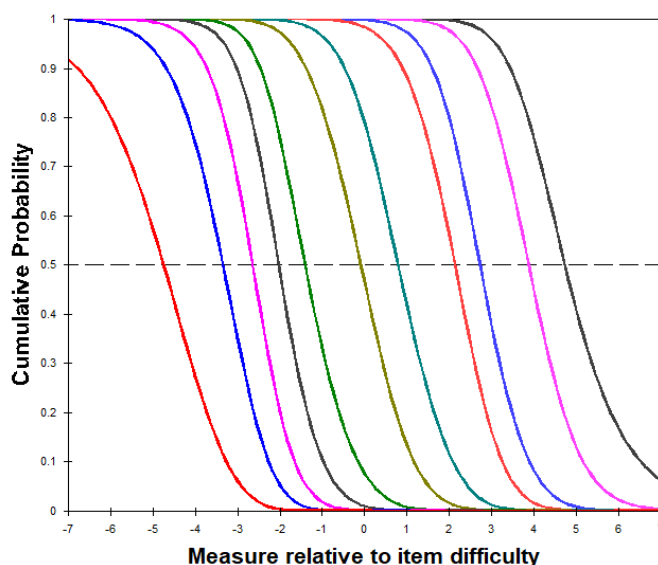
The relatively even spacing tentatively confirms O'Fallon's hypothesis that the *STAGES model cuts the developmental space into equal-spaced units*. O'Fallon proposed that one benefit of a theory-based (i.e., explanatory vs. more descriptive and data-driven) model was that it offered a consistent "ruler" against which to measure the progression of stages, which is resilient to data collected in future studies from different populations. This hypothesis is supported here.

**Category boundaries and step uniformity.** The PCC (Figure 11) also shows that the separations between values (the curve intersections or Rasch-Andrich thresholds) is relatively even over the range, which represents a primary predication by O'Fallon.

A better method of illustrating the category boundaries regards the category demarcation graph shown in Figure 12, showing the relationship between the STAGES score for each item vs. the Rasch model measure of the latent variable. The relationship is quite linear within the normal



**Figure 12.** Stages score vs. Latent Variable Measurement.



**Figure 13.** Cumulative probability curves.

which has some levels which are more highly represented. 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 tends to have a normal-shaped distribution). It is however pertinent to note that the "maximum strata" for our system is calculated at 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 the less well represented levels).

### Summary/fit table and

**discrimination strata.** Figure 14 shows the summary table with fit statistics from the WINSTEP output. INFIT and OUTFIT statistics assess outliers and flag possible problems. The fit statistics identifying mis-fitting (infit and outfit) items is .99, which is in the "very good" range (i.e., very low misfit means – values less than 2 are considered acceptable). The reliability of the items is **excellent** at .95 and is analogous to the excellent Cronbach's alpha metric previously noted for the test.

The separation index of persons at 5.52 is noteworthy. Through a transformation formula, this provides an indication of the number of "strata" or levels that the test resolves. (This 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  $(4S-1)/3$  meaning that  $(4*5.52-1)/3 = 7.0$  strata. This is analogous to the estimate above with the PPC curve that the test cleanly differentiates the 6 person perspectives (7 levels according to this calculation) and has more difficulty resolving all 12 levels as clearly.

The strong differentiation of 6 (or 7) levels holds for our particular dataset,

<b>PERSON</b>	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	
<b>ITEM</b>	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 14. Winsteps Summary Table.

**Rasch/IRT analysis summary.** Whether examining Winsteps (Rasch) or R (IRT) outputs, our consultants were unanimous that the *STAGES inventory and scale is psychometrically sound across its entire range* (with more certainty about this for the levels inside the extreme ratings). Its strength had been verified in the past by repeatedly high Cronbach's alpha measurements of internal consistency ("reliability"), but the IRT and Rasch analysis contribute additional depth to this finding.

The test would be strong psychometrically at 18 or even 10 items. All items on the general protocol are strong, but the stem "I am" stands out as fitting less with the model compared to the others. The 12 stages defined by the model are relatively evenly spaced along the spectrum, supporting O'Fallon's conjecture that the STAGES model cuts the range into equal slices.

The test easily differentiates the 6 person perspectives and has more difficulty differentiating between levels at the granularity of 12 levels. We will investigate how the scoring manual or procedure might provide a more refined differentiation between the first and second halves (the actively and passively oriented) of each person perspective. According to O'Fallon (personal communication), this makes sense within the theoretical model. At the X.0 or passively oriented levels, the meaning-making system of the prior X.5 (actively oriented) level is being deconstructed and a new worldview is emerging. This process might arise in language with more variation and less precision in textual measurement.

## A New Aggregation Method – The Core Stage

We have outlined many problems with the ogive cutoff method of aggregating item scores to produce the TPR. IRT (and Rasch) analysis ignores the ogive formula and assumes that the latent variable is a simple or weighted combination of the item scores, which is performed for most psychometric assessments. This is analogous (though not exactly the same as) the less-used TWS score for a survey.

Regarding the IRT and Rasch analysis, our consultants agreed that (1) the test is valid and strong examined at the item level and (2) the ogive cutoff system was problematic. We still want our developmental scoring to be compatible with the expectation that, for a projective test, a person's true "center of gravity" will be evidenced from the higher scores, and that how the lower

scores are spread out does not affect this summary measurement. To achieve this, we must discard some assumptions made in IRT, as none of our consultants were able to identify any psychometric evaluations of aggregation methods for projective tests (or tests that focused on only the top items). Thus, in this respect the SCT remains somewhat an anomaly within the field of psychometrics, which is why Loevinger had to develop an alternative method (the ogive method).

We had two primary goals for designing a new aggregate method, which we ended up calling the "Core Stage."<sup>15</sup> First, we wanted to avoid the problems of the ogive cutoff method and use something more straightforward and with fewer arbitrary parameters. Second, we wanted to minimize the differences between the new and old scores so that we would not have to alter the established interpretation of each level (i.e., to remain consistent with how a "3.5/Achiever" total score is understood in the field). Once a method was determined, we wanted to scale it to appear as close to the prior ogive method as possible (based on the RMS error of the Core Stage vs. the ogive score).

We considered two methods: The first regarded taking the average of the top X scores, where X was to be experimentally determined. 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 that systematically increased with higher scores, where the power was to be experimentally determined. By "experimentally," we mean iterating over different values to identify which come closest to the old ogive score. We had 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 appeared favorable for surveys at all levels.

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.)

The results are shown in Table 2. We compare each new method with the old ogive score by taking the root mean square error over the entire database (n=740). "Ave of all" is comparable with the TWS method (which regards the sum of the scores) and shows a baseline RMSE of 0.55. Note that because the scores increase by 0.5 per level, an RMSE of 0.5 represents one full level. The lowest RMSE for the TopX method was 0.22 at Top10 and Top12. When taking the average of the sum of the score raised to powers, the power with the lowest RMSE was  $\text{Score}^8$ . It appeared that this method did not produce errors as low as the TopX method and was also more difficult to conceptually resolve with our understanding of developmental progress. (We also tried combining the two methods, which did not yield a lower RMSE.)

---

<sup>15</sup>This term may be revised when a better one is found.

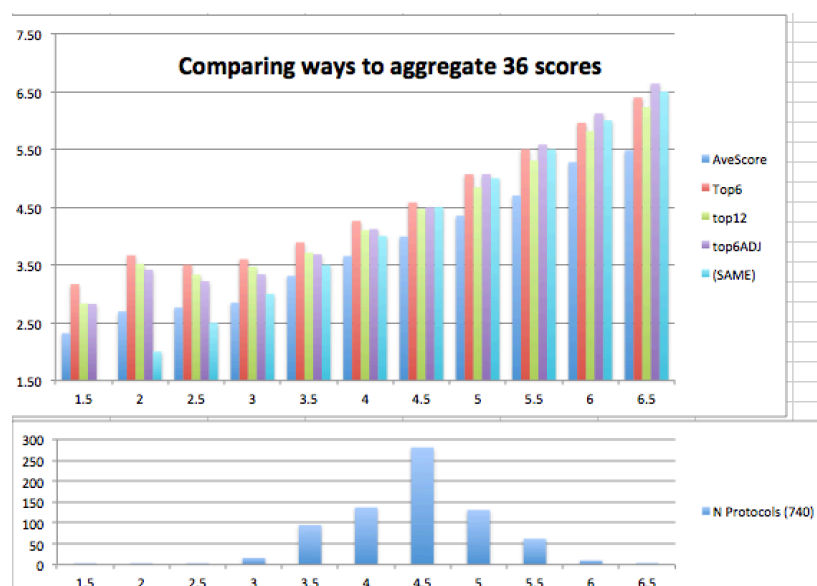
**Table 2.** RMS Error of alternative aggregate methods.

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
aveTop10Score^2	0.24

The RMS error represents an average error over all surveys, however we wanted to more closely examine 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 interpretation of low-level scores. 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 15 shows histograms displaying the scores of the new methods (Y axis) vs. the old ogive score (x axis). The bar marked "SAME" provides comparison to illustrate the "straight line" one would achieve if a new method gave the exact same score as the ogive method. Although "top 10" provided the best results, we decided to use either Top6 or Top12, which had RMSE's practically as good since they represent exactly 1/6th or 1/3rd of the 36 stems. Top6 always performs better than Top12, and both always perform better than the average, so we choose the method of (average of the) Top6 (or top 6th for surveys with other than 36 items) for the new Core Stage.

Next, to bring the Core Stages even closer to the old ogive scores, we performed a linear transformation of the Top6 score. Linear fitting showed that a formula of  $(1.2 \times \text{AveOfTop6}) - 1$  would yield a score even closer to the old ogive. We call this the "adjusted" Top6, which is also shown in Figure 14. The adjusted top6 performs better for most of the subtle and metaware levels, and it is unimportant that it does not perform as well for the Concrete tier.



**Figure 15.** Comparing aggregation methods across all levels.

- It yields results that are usually 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 can thus be useful for scoring children (which O'Fallon has begun doing).
- It is not based on a suspicious or erroneous application of Bayes' theorem.
- It does not force scores into categories but allows fractional values.
- It is not sensitive to small differences in how stems are scored (which occurs in the ogive method near the cutoffs).
- The new Core Stage rating makes it easier to produce a total score for protocols of any length (one does not have to worry about inventing new cutoffs).
- It does not contain a large number of arbitrary parameters (such as the 12 ogive cutoffs and 12 TWS multipliers and the other Bayesian parameters). The primary ad-hoc

<sup>16</sup> When N is not divisible by 6: Where  $36/N = I(\text{nteger}).R(\text{emainder})$  (e.g 2.667), Ave\_top\_6th is the weighted mean where the top I are given weight 1, and the I+1th score is given a weight of R, the rest given weight 0.

parameter is the "6" (or 1/6th) by taking the top 6 scores. Our analysis indicates that outcomes are not overly sensitive to this parameter (i.e., 5 to 10 behaves similarly), however tuning this parameter remains an open question. In addition, the specific formula for aligning the Core Stage with prior ogive values is dependent on the dataset used and may need to be adjusted in the future.

This Core Stage measure has been added to the STAGES assessment software (on both the "Training Platform" and the "Production Platform"). Both the new Core Stage (and additional metrics discussed below) and the prior ogive and TWS metrics are currently shown. We are in a period of training scorers to use and understand the new method, however the old method will be used for at least another year for reporting results to clients.

For several months, O'Fallon has been observing that the Core Stages closely match her intuitions about developmental stages, are close to the old ogive score, and function 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 Core Stage: 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 Core Stage minus the Bottom Score, and indicates the range of values in the survey.

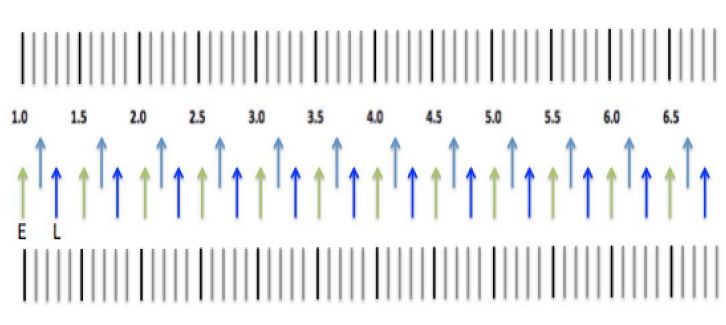
**Early and Late stages.** Prior to this work and continuing after it, advanced scorers use an additional set of text analysis instructions to determine a sub-level rating of early, middle, or late within each level (for each stem). Although this refinement of the item score is not included in the ogive (or TWS) calculation, it can be considered when the total survey score is tweaked based on global considerations.

The new Core Stage allows for fractional values, which provides another indication of early/middle/late sub-levels within a stage. We still need to further consider how these two methods of sub-stage scoring (i.e., Core Stage fractional level vs. text analysis rules for sub-level) will coordinate with each other. These two methods will not necessarily coincide, and it remains to be determined how they will be resolved.

In addition, the question arises RE where along the fractional scale we call "early" and "late." For example, *early* 3<sup>rd</sup>-person perspective could be just *below* "3.0" as in "2.9 or 3.0" or it could be just at and above 3.0 as in "3.0 and 3.1." We have decided tentatively to use the convention shown in Table 3 and illustrated in Figure 16.

**Table 3.** Decimal fractions vs. early and late for stage X.

X.0 early	X.5 early
X.1 mid	X.6 mid
X.2 mid	X.7 mid
X.3 late	X.8 late
X.4 transition between	X.9 transition between



**Figure 16.** Early and Late stages, illustrated  
(Early in green, late in dark blue, middle in light blue).

As indicated, the fractional Core Scores will be rounded to the nearest single decimal value. In reports that classify scores at the granularity of each developmental level, as if "rounding" to a developmental level, we will round the X.4 and X.9 fractions up and the other fractions down. For example. 2.9, 3.0, 3.1, 3.2, 3.3 will be rounded to "3.0" and 3.4, 3.5, 3.6, 3.7, 3.8 will be rounded to "3.5."

Although this has not yet attempted, it would be possible to adjust the item scores based on the sub-level text analysis of the stem completions. The Core Stage could be adjusted in the following manner if early or late indicators were found:

- for early, add 0
- for unspecified or middle, add 0.15
- for late, add 0.35

This corresponds with the method shown in Table 3 and Figure 16.

## Conclusions and Future Work

This research has increased our confidence in the STAGES theory, model, and scoring methodology and has contributed some discoveries that arguably, could apply to other SCT frameworks as well. The SCT method of measuring ego development generally stood up well under the lens of IRT and Rasch analysis.

We have also produced a new method for calculating the aggregate (total) score over the 36 stems (or fewer), which we call the Core Stage. The Core Stage has many desirable properties, as noted in the prior section. The STAGES framework will be gradually shifting over to this new method, however we remain agnostic regarding whether other SCT frameworks should adopt this method (or any alternative to the ogive method). The ogive method has been performing well enough over the years, represents the default standard, and the effort to adopt a new method may well outweigh any benefits. The benefits of changing are greater for the STAGES model vs. other SCT variations, because others use a fixed set of stems (sometimes splitting the survey in half), while the STAGES allows for creating valid surveys with new stems of any length, which would require ongoing alteration of cutoff values, which can only be performed in arbitrary ways. In addition, automated computer scoring exists for the STAGES model (see [www.stagelens.com](http://www.stagelens.com)), which means the SCT will become more practical to use for large-scale research where the validity of methods is critical.<sup>17</sup> (The companies using the MAP, LGP, and

<sup>17</sup> This automated scoring is not as accurate as human scoring, but is accurate enough for statistically valid studies of groups of 20 or more.

GLP are exclusively focusing their assessment work on individual assessments and the quality of their coaching and consultation with individuals and teams, as far as we know.)

**Study limitations.** Limitations of our research methodology and results include the fact that the sample population is on average higher than the general population. Although IRT analysis is expected to be fairly resistant to such skew effects, the results may be more robust if we obtain more scores from the 2.5, 3.0, and 3.5 levels, which we will attempt to do. The analysis was limited to the STAGES general protocol and to data collected before mid-2018. A larger dataset is now available and might be used to update some results.

Although the Core Stage method does not contain a large number of arbitrary parameters (such as the 12 ogive cutoffs, the 12 TWS multipliers, and the other Bayesian parameters), it does contain some arbitrary parameters, specifically the "6" (or 1/6th) when taking the top 6 scores. Our analysis indicates that outcomes are not overly sensitive to this parameter, and the specific formula for aligning the Core Stage with prior ogive values is dependent on the dataset used and may need to be adjusted in the future.

**Future research.** Future research could further explore:

- Given the predominance of 4.5 surveys, it would be useful to acquire more data in the database for lower stages.
- RE the "strata" metric and weaker ability to differentiate all 12 levels, it might be possible to modify the coding manual to better differentiate early from late stages (and active vs passive), one method for which would be to compare surveys with Core Stages around "X.5" with those at X.0 above and below them and seek additional principles that would differentiate the X.5's.
- We could more closely examine the differences between stem items regarding their average difficulty and standard deviations and which levels they better perform at differentiating.
- We could more closely examine what occurs with the "I am" stems and determine whether additional stems could be invented that target the same sub-capacity.
- We could investigate the dip in the number of stems scored at 4.0 over all items.

## References

- Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage Publications.
- Bond, T., G. (1994). Piaget and measurement II: Empirical validation of the Piagetian model. *Archives de Psychologie*, 63, 155-185.
- 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.
- Linacre, J. M. (2018). *A User's Guide to WINSTEPS and MINISTEP Rasch-Model Computer Programs* (version 4.3.0). Available at winsteps.com.
- 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 of a Sentence Completion Test, Vol. 1*. San Francisco: Jossey-Bass.
- Müller, U., Sokol, B., and Overton, W. F. (1999). Developmental sequences in class reasoning and propositional reasoning. *Journal of Experimental Child Psychology*, 74(2), 69-106.
- 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. (to be published, 2020). The Validation of a New Scoring Method for Assessing Ego Development Based on Three Dimensions of Language.
- O'Fallon, T. (2011). STAGES: Growing up is Waking up – Interpenetrating Quadrants, States and Structures. Available at [www.pacificintegral.com](http://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.

- Redmore, C. (1976). Susceptibility to faking of a sentence completion test of ego development. *Journal of Personality Assessment*, 40(6), 607-616.
- 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](http://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.
- Wilber, K. (1995), *Sex, Ecology, Spirituality: The Spirit of Evolution*. Boston: Shambhala.

## Appendices

### Appendix 1: TWS calculation and ogive cutoffs

The original STAGES TWS formula is an extension of that used by Cook-Greuter and Loevinger and 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

---

Table 4 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, and 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 regards 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 tested begin with the highest (6.5), move down to 3.0, and then begin 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 represents the final or "default" value in the procedure.<sup>18</sup>

Cook-Greuter used the cutoff of 4 for the 2 levels above Strategist (4.5), and STAGES duplicated that approach by using 4 for all stages above 4.5.

---

<sup>18</sup> 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.

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

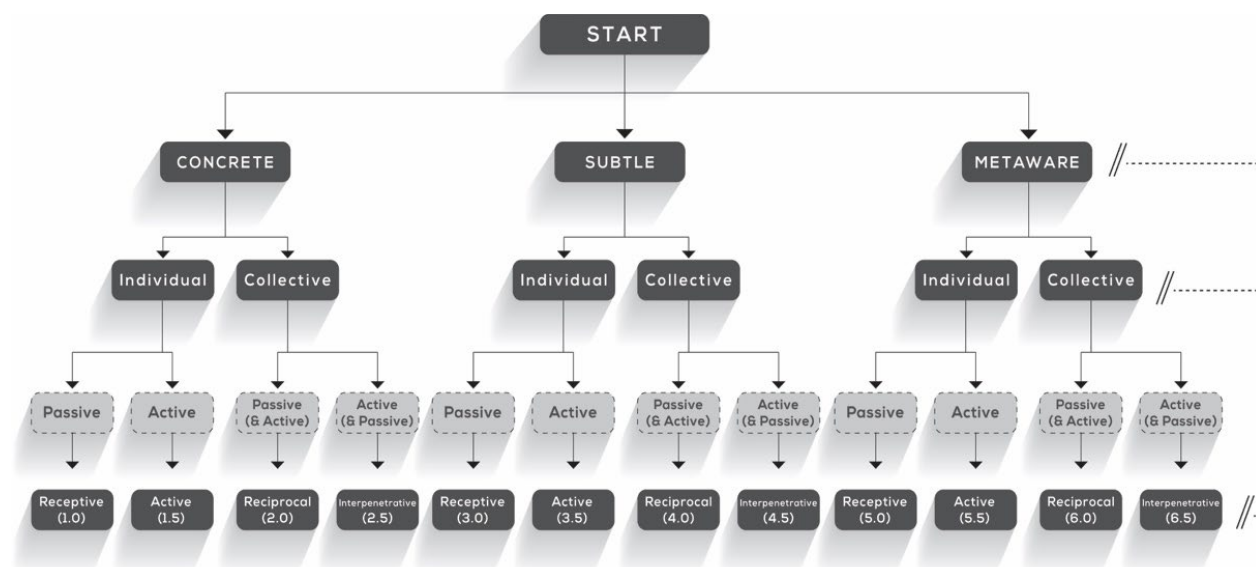
<i>(Do steps in order and stop where true)</i> If there are:			Assigned stage
4	or more rated at	6.5 or higher	6.5
4	or more rated at	6.0 or higher	6.0
4	or more rated at	5.5 or higher	5.5
4	or more rated at	5.0 or higher	5.0
6	or more rated at	4.5 or higher	4.5
9	or more rated at	4.0 or higher	4.0
14	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.

## Appendix 2: Stages Model Figures

Below are three descriptions of the 12 levels of the STAGES model. Figure 17 shows how the scoring process determines the level using three questions, Figure 18 shows the same information in a different graphical format, and Table 5 shows how the names of STAGES levels correspond to names given in related developmental theories.



**Figure 17.** Diagram of stage assigned based on the responses to three questions.

<i>Person-Perspective</i>	<i>Stage</i>	<i>Tier</i>	<i>I/C</i>	<i>A/P</i>			
<b>STAGES Model</b>	6 <sup>th</sup>	Metaware	C	A	Interpentr		
			C	P	Reciprocl		
	5 <sup>th</sup>		I	A	Active		
			I	P	Receptive		
	4 <sup>th</sup>		Subtle	C	A	Interpentr	
				C	P	Reciprocl	
	3 <sup>rd</sup>			I	A	Active	
				I	P	Receptive	
	2 <sup>nd</sup>			Concrete	C	A	Interpentr
					C	P	Reciprocl
1 <sup>st</sup>	I	A			Active		
	I	P			Receptive		

**Figure 18.** STAGES Tiers & Repeating Principles  
Asterisks (\*) show levels added in STAGES vs MAP model.

O'Fallon has speculated a fourth tier in the STAGES model, continuing the repeating patterns observed in the first three tiers into 7.0, 7.5, 8.0, and 8.5. This fourth tier is based on her reading of Aurobindo's description of these rare stages of development.

Below is a table showing correspondences between five developmental models gleaned from sources in the literature. Definitions of levels of different models do not usually exactly align but share sufficient characteristics that they are comparable.

**Table 5.** Correspondences of Levels in Five Developmental Models.

<b>O'Fallon</b>	<b>Loevinger</b>	<b>Cook-Greuter</b>	<b>Torbert</b>	<b>Kegan</b>
6.5 (Illumined)				
6.0 (Universal)		Unitive	Ironist	
5.5 (Transpersonal)				
5.0 (Construct Aware)	Integrated e9	Construct-aware / Ego-aware	Alchemist/Magician	
4.5 (Strategist)	Autonomous e8 (i5)	Autonomous, Self-actualizing	Strategist	Self-transforming/Inter-individual (5th)
4.0 (Individualist)	Individualistic e7 (i4/5)	Individualist, Self-Questioning	Individualist	
3.5 (Achiever)	Conscientious e6 (i4)	Conscientious, Self-Determining	Achiever	Self-authoring/Institutional (4th)
3.0 (Expert/Specialist)	Self-aware e5 (i3/4)	Self-conscious, Skill-Centric	Expert/Technician	
2.5 (Diplomat/Conformist)	Conformist e4 (i3)	Conformist, Group-Centric	Diplomat	Socialized/Interpersonal (3rd)
2.0 (Rule oriented)	(Delta/3)			
1.5 (Opportunist/Egocentric)	Self-protective e3 (i 2/3, opportunist, delta)	Self-defensive, Self-Centric	Opportunist	Instrumental/Imperial (2nd)
1.0 (Impulsive)	Impulsive e2 (i2)	Impulsive	Impulsive	Impulsive (1st Order)
	Pre-social e1 (i1)			

Most scholars agree that there is substantial overlap and that the various theories are more or less discussing the same thing. There are however differences as well, and these become important as one tries to compare one theory to another or assessments made within one model to assessments made within another. Although there are differences, the similarities and alignments are actually quite striking since many were independently developed, which supports the overall validity of the developmental approach.

### Appendix 3: Item response theory and Rasch analysis

IRT (also known as latent trait theory, strong true score theory, or modern cognitive test theory) represents a main school/toolbox 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 (called the "latent variable" since it cannot be directly measured). Rasch analysis is usually considered a subset of IRT that, for theoretical reasons, is limited to "one-parameter" modeling (IRT includes a family of 1-, 2-, and 3-parameter models).

Whereas they are well-known in psychometric circles, Rasch's (1980) models for measurement have been employed by developmental psychologists only recently (Andrich & Constable, 1984; Bond, 1994; Dawson, 1998, 2000; Müller, Sokol, & Overton, 1999).

Although IRT does not address the question of how to aggregate item scores for a projective test, we performed an IRT analysis of the STAGES database of scored protocols to establish a firm foundation for whichever aggregation method we chose to replace the ogive cutoff method.

**Item Response Theory.** The table below illustrates key differences between classical test theory and IRT (including Rasch analysis). (The Wikipedia entry on IRT also provides a useful overview.)

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

**Figure 19.** Classical Test Theory vs. Item Response Theory.

Item response theory creates a single scale 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), which allows item difficulties and person abilities to be meaningfully compared. For example, a value of "3.2" means the same thing regarding the latent variable when used to discuss the skill level of an individual, the difficulty level of an item, or the overall score on a test. Item response theory can calculate the difficulty and information power of each test item at each trait level.<sup>19</sup> For example, an item or an entire test can be found to reliably differentiate weak from moderate ability but can be poor in differentiating between moderate and advanced abilities.

One thing that distinguishes IRT from "classical" test theories is that it does not assume that each item is equally difficult, meaning it calculates an item difficulty for each item. Item response theory 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. Item response theory analysis begins with a sufficiently large set of test data and, using that assumption, iteratively works backwards to calculate (1) the difficulties of each item and (2) the overall skill of each test-taker (this type of method was infeasible before computers, which is one reason it was less popular when Loevinger began her research).

Item response theory converts the measurement (latent) scale of test into a standardized form. By using "logistic" scaling, the 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 cover a specific range as in a 1-5 Likert scale or a 0-100% correct score for a math problem); and (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). Item response theory models posit that the difference between an item's difficulty and a person's ability reflects the probability of a person succeeding at 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 "polytomous" IRT methods were used here, meaning those applying to non-binary measurements.)

Item response theory 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)? For more information, see the figure in Appendix 4 "IRT vs. Classical test theory."

- calculate total test scores that take into account variations in the strength (reliability or information power) of each item;
- determine the reliability of items and of the test as a whole, including the information power of items (or the test) to assess each part of the scale;

---

<sup>19</sup> 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.

- find outliers or misfits in test items, indicating that they do not function as expected and should be changed or removed;
- identify outliers in the test-taking subjects, such as those who answer typically difficult items correctly but answer typically easy items incorrectly;
- estimate the similarity of the gaps between ordinal trait categories and the resolution or number of such categories that a test can differentiate;
- test for the uni-dimensionality of the construct regarding whether the test as a whole measures a single latent variable.

**Rasch Analysis.** Rasch analysis (Rasch, 1980; Andrich, 1988; Bond & Fox, 2001) is usually considered a subset of IRT, although many Rasch proponents maintain that its underlying philosophy of measurement theory is distinct. Item response theory includes 1-, 2-, and 3-parameter models, while Rasch proponents advocate for no more than 1 parameter. "Parameters" regard numbers that characterize each item differently, where the first and primary parameter is the item difficulty and 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 or higher 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 regards 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, which means that the test item should be removed or revised. In contrast, the more standard IRT method adopts a purely statistical modeling approach that uses as many parameters as needed to find a model that best fits all the data. The details of this controversy exceed our scope, but our consultants have indicated that for our dataset, the difference does not matter and that the 1-parameter analysis compatible with Rasch assumptions is sufficient.

Commons and Pekker (2009, p. 13-14) summarize the Rasch model by 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."